Protocolos de Difusão Periódica de Vídeo sob Limitação de Banda Passante

Rogério Makiyama Zafalão

Dissertação de Mestrado

Instituto de Computação
Universidade Estadual de Campinas

Protocolos de Difusão Periódica de Vídeo sob Limitação de Banda Passante

Rogério Makiyama Zafalão[1]

Fevereiro de 2004

Banca Examinadora:

- Prof. Dr. Nelson Luís Saldanha da Fonseca (Orientador)
- Prof. Dr. Edmundo Roberto Mauro Madeira
  Instituto de Computação, Unicamp
- Prof. Dr. Edson dos Santos Moreira
  Instituto de Ciências Matemáticas e de Computação - ICMC-SC, USP
- Prof. Dr. Cid Carvalho de Souza
  Instituto de Computação, Unicamp

[1]**Trabalho parcialmente financiado pela CAPES (convênio DS-108/00)**

Protocolos de Difusão Periódica de Vídeo sob Limitação de Banda Passante

Este exemplar corresponde à redação final da Dissertação devidamente corrigida e defendida por Rogério Makiyama Zafalão e aprovada pela Banca Examinadora.

Campinas, 27 de fevereiro de 2004.

Prof. Dr. Nelson Luís Saldanha da Fonseca
(Orientador)

Dissertação apresentada ao Instituto de Computação, UNICAMP, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Agradecimentos

- Ao meu orientador, Nelson, pelo comprometimento e pela paciência, um agradecimento especial pela oportunidade de sua orientação neste trabalho;

- à toda a família, pelo amor, confiança e apoio incondicionais;

- à Fábia, pelo amor, carinho e compreensão;

- aos colegas do IC, pela ótima companhia;

- à CAPES, pelo apoio financeiro deste trabalho;

- a Deus.

Resumo

Sistemas de Vídeo sob Demanda possibilitam ao usuário a escolha de vídeos para exibição dentre uma vasta coleção. Como a transmissão de um fluxo de vídeo demanda uma grande quantidade de bande passante, estratégias de compartilhamento da transmissão de fluxos de vídeo são utilizadas para reduzir esta demanda. Dentre estas técnicas, os protocolos baseados em difusão periódica são indicados para a transmissão dos vídeos mais requisitados, uma vez que estes utilizam largura de banda constante independente do número de usuários. Entretanto, os protocolos mais eficientes não levam em consideração limitações de banda passante existentes no cliente.

Nesta dissertação, dois protocolos de difusão periódica otimamente estruturados são estendidos de forma a permitir que clientes sujeitos a limitações de largura de banda possam utilizar serviços de Vídeo sob Demanda baseados em difusão periódica.

Palavras-chave: Multimídia, Vídeo sob Demanda, Protocolos de Difusão Periódica, PDPH-LBU, GEBB-LBU

Abstract

Video on Demand (VoD) services allow users to watch movies of their choice among a wide collection. As video transmition requires a huge amount of bandwidth, stream sharing techniques has been developed to reduce the bandwidth requirements. Among these techniques, periodic broadcasting protocols are indicated to transmit most frequently requested videos, since they require a constant amount of bandwidth. However, these protocolos do not consider users with limited bandwidth.

In this dissertation two new protocols are introduced, the Polyharmonic Broadcasting with Limited User Bandwidth (PHB-LUB) and the Greedy Equal-Bandwidth Broadcasting with Limited User Bandwidth (GEBB-LUB).

Keywords: Multimedia, Video on Demand, Periodic Broadcasting Protocols, PHB-LUB, GEBB-LUB

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

Vídeo sob Demanda (*Video-on-Demand*, ou VoD) compreende um conjunto de aplicações que possibilitam a seleção e exibição de uma grande variedade de vídeos através de uma rede de computadores. Como exemplo, pode-se citar o ensino a distância, as bibliotecas digitais, a distribuição de notícias, os jogos, o *home shopping* e vídeos para entretenimento (videoclipes e filmes). Cabe ainda destacar o *Movie on Demand* (MoD), um serviço específico dentro de VoD, onde os objetos de vídeo são exclusivamente filmes, alcançando um público que atualmente utiliza serviços de *pay-per-view* (PPV) e locação de vídeos.

Como vídeos são objetos que requerem grande quantidade de banda passante para exibição, a largura de banda torna-se um fator limitante no fornecimento de serviços VoD. Assim, é extremamente importante a utilização racional da banda passante na transmissão de fluxos de vídeo. Vídeos "populares"[1] possibilitam que vários clientes recebam todo o fluxo (ou parte dele) através de uma única transmissão via difusão (*broadcast*) ou difusão seletiva (*multicast*).

Para o correto funcionamento de VoD, é necessária a utilização eficiente de recursos, tanto no lado do cliente quanto no lado do servidor, tais como: largura de banda, capacidade de armazenamento e taxa de transferência de E/S. A grande maioria dos protocolos de difusão periódica assume que os clientes não estão sujeitos a limitação de recursos, o que não é realista.

O presente trabalho introduz extensões dos protocolos de Difusão Poliharmônica e o Protocolo de Difusão Guloso com Canais de Mesma Largura de Banda (*Greedy Equal Bandwidth Broadcasting* - GEBB), ambos otimamente estruturados (*optimally-structured*), de forma que estes possam ser utilizados sob limitação de banda passante no cliente. Estas

[1]Também chamado de *hot videos*, correspondem aos vídeos com maior freqüência de requisição.

extensões são baseados na Teoria de Otimização para determinar a configuração ótima dos parâmetros dos protocolos, de forma que o servidor utilize a menor quantidade de banda passante possível.

A presente dissertação está organizada em 6 capítulos, dispostos da seguinte forma:

No Capítulo 2, faz-se uma revisão literária sobre o serviço Vídeo sob Demanda.

No Capítulo 3, faz-se uma breve introdução de algoritmos genéticos.

No Capítulo 4, são detalhados os protocolos de VoD que se utilizam de difusão periódica.

No Capítulo 5 introduz-se a extensão dos protocolos otimamente estruturados para usuários com limitação de banda passante. São apresentadas neste capítulo dois novos protocolos, uma extensão do Protocolo de Difusão Poliharmônica e outra ao o protocolo GEBB.

Por fim, o último (Capítulo 6) traz um resumo dos resultados obtidos, além de considerações finais e sugestão de trabalhos futuros nesta área.

Capítulo 2

Conceitos de Sistemas de Vídeo sob Demanda

Vídeo sob Demanda é considerada a "*killer application*" para os novos serviços de comunicação. VoD permite que um usuário, no seu horário de preferência, selecione vídeos para exibição dentre uma coleção (e.g. filmes, aulas e vídeos educacionais, notícias e jogos interativos) através de uma rede de computadores.

Este capítulo apresenta vários conceitos e protocolos utilizados na área de VoD. A Seção 2.1 apresenta a arquitetura utilizada em sistemas de VoD, bem como seus componentes principais: servidor, rede de distribuição e o equipamento do usuário.

Na Seção 2.2, são apresentadas as técnicas gerais para redução de banda passante, a motivação para a utilização destas técnicas e, por fim, a notação utilizada nas seções seguintes para explanação dos diversos protocolos de redução de demanda de largura de banda.

A Seção 2.3 aborda técnicas de redução de banda passante que se utilizam de difusão seletiva. Particularmente, as que utilizam várias técnicas em conjunto, como *Batching* e *PiggyBacking*, ou Mcache.

2.1 Arquitetura de um Sistema de Vídeo sob Demanda

Nesta seção, são apresentados os principais componentes de uma arquitetura convencional de VoD: servidor, rede de distribuição e o equipamento do usuário. Também são apresentados alguns conceitos necessários ao desenvolvimento do presente trabalho, bem como

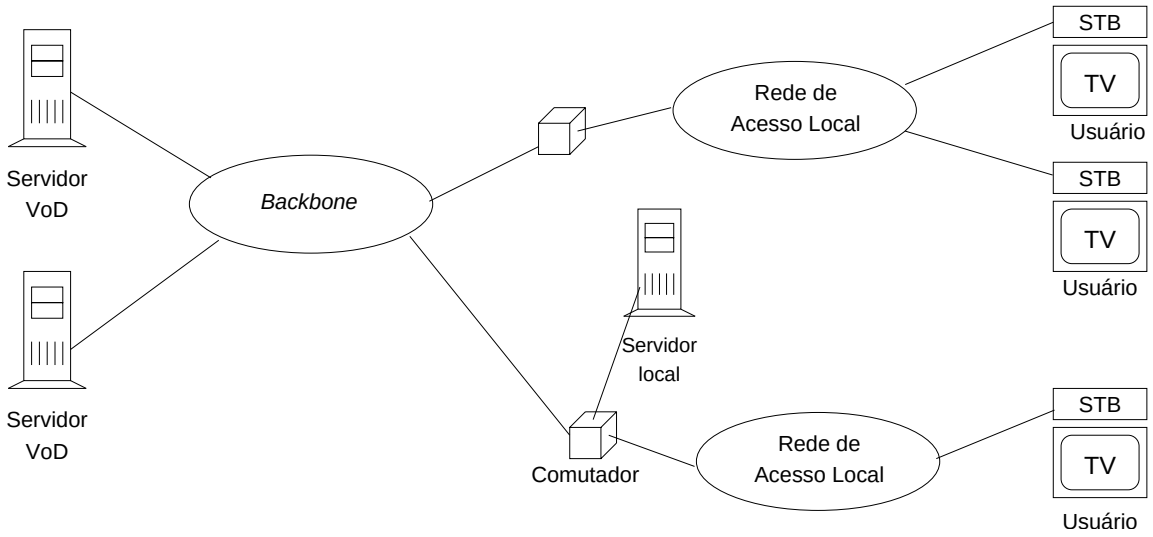

Figura 2.1: Arquitetura de um sistema VoD

também as classificações principais utilizadas em VoD.

2.1.1 Descrição da Arquitetura

Um sistema típico de VoD [46, 12, 20] possui um conjunto de servidores de vídeo e clientes distribuídos geograficamente, conectados através de uma rede de distribuição de alta velocidade. Propriedades desejáveis para uma arquitetura de VoD incluem baixa latência, escalabilidade e flexibilidade no que se refere a restrições de recursos de clientes. A Figura 2.1 ilustra a visão geral de um sistema de vídeo sob demanda.

Servidores locais (*spooling servers*) atuando como *proxies* [46, 48, 20] nesta arquitetura são tidos como elementos importantes para o desempenho dos sistemas de VoD, especialmente quando utilizados em conjunto com a técnica de *prefix caching*[40, 45].

2.1.2 Servidores de Vídeo

Um servidor de vídeo compreende uma hierarquia de armazenamento dos vídeos, controladores de fluxos, barramento e processador.

Hierarquia de Armazenamento

O sistema de armazenamento do servidor precisa suportar um grande número de usuários simultaneamente, assim como o controle de *hardware* e *software* para o tratamento de

requisições, localização de vídeos e transferência de dados entre dispositivos na hierarquia de armazenamento [46, 42, 29]. Os recursos de armazenamento são vistos como uma hierarquia cujos dispositivos são classificados pela capacidade de armazenamento e tempo de acesso, composta por:

Armazenamento terciário Compreendem dispositivos de armazenamento que possuem baixo custo, grande capacidade de armazenamento, tempo de acesso elevado e baixa taxa de transferência. Como exemplo, pode-se citar mídias como CD-ROMs, DVDs e discos óticos organizados em *jukeboxes*. *Jukebox* é um grande agrupamento de mídias dispostas em estantes e acessadas por dispositivos robóticos, cujos braços mecânicos substituem o último disco utilizado pelo que contém o vídeo a ser acessado.

Armazenamento secundário Consiste em tecnologia de armazenamento ótico de alta velocidade ou sistemas RAID (*Redundant Arrays of Inexpensive Disks*).

Armazenamento primário Discos convencionais, organizados ou não na forma de RAID.

Memória RAM e *cache* A memória RAM é utilizada num servidor de vídeo (além do seu uso comum) para transmissão de fluxos individuais ou difusão.

2.1.3 Rede de Distribuição

A rede de distribuição consiste em um *backbone* interligando o servidor de vídeo à rede de acesso, e compreende todo o conjunto de *switches* e cabeamento necessários para sua interligação.

Backbone

ATM (*Asynchronous Transfer Mode*) é o padrão para as Redes Digitas de Serviços Integrados de Faixa Larga (RDSI-FL) e é utilizado em vários *backbones* na Internet. Além de ATM, é também utilizado IP tanto sobre WDM como também sobre SONET (*Synchronous Optical Network*). Na Internet, tentativas de controle de tráfego eficientes como MPLS, IntServ e DiffServ ainda são estudadas.

Rede de Acesso

As redes locais de acesso são responsáveis pela ligação entre o *backbone* e o equipamento do usuário. Na Internet, por exemplo, correspondem aos *ISP - Internet Service Providers*.

As soluções mais comuns são listadas a seguir:

ADSL (*Asymmetric Digital Subscriber Line*): Utiliza o cabeamento de telefonia já existente, oferecendo um canal para transmissão de dados a até 1,536 Mbps, um canal para sinais de controle a 16 Kbps e mais um canal analógico para telefonia, operando a uma freqüência de 4 kHz.

HFC (*Hybrid Fiber/Coax*): Esta solução propõe a utilização de um cabo coaxial compartilhado, interligando diversos usuários a um segmento de fibra ótica ligado ao *backbone*. Suas transmissões podem ocorrer de forma analógica (TV a cabo) ou digital (VoD). Devido ao acesso compartilhado, é necessário algum mecanismo de acesso múltiplo, como TDMA(*Time Division Multiple Access*) ou FDMA(*Frequency Division Multiple Access*). O HFC fornece uma grande largura de banda, suportam centenas de conexões simultâneas e permite a utilização de estratégias de compartilhamento de fluxo, trazendo benefícios para a implementação de protocolos baseados em difusão seletiva ou difusão periódica.

2.1.4 *Set-Top-Box*

O equipamento utilizado por um usuário final inclui a *set-top-box* (STB) conectada a um aparelho de TV, vídeo cassete ou PVR (*Personal Video Recorder*). A *set-top-box* permite aos usuários selecionar o serviço ou vídeo de sua preferência, possui uma unidade de descompressão para decodificar a seleção realizada, geralmente codificada em MPEG-1 ou MPEG-2, e enviá-la ao dispositivo de TV.

Técnicas descritas nas Seções 2.3 e 2.4 (difusão seletiva e difusão periódica, respectivamente) requerem uma memória auxiliar, ou *cache*, para potencializar a redução de banda passante de fluxos de vídeo.

2.1.5 Conceitos Básicos Utilizados em VoD

A seguir, são apresentados alguns termos e conceitos utilizados com maior freqüência na literatura relacionada a VoD, e que serão utilizados nesta dissertação.

CBR (*Constant Bit Rate*) Fluxo de vídeo cuja principal característica é a taxa de exibição constante ou aproximadamente constante.

VBR (*Variable Bit Rate*) Característica de um fluxo de vídeo onde sua taxa de exibição não seja constante, como decorrência da compactação, como por exemplo, em vídeos codificados em MPEG-2, MPEG-4.

Codificação MPEG MPEG, ou *Motion Picture Expert Group*, é uma família de padrões internacionais definido para a compressão e transmissão de áudio e imagens em movimento. O MPEG explora o fato de um fluxo de vídeo possuir altos níveis de correlação entre suas imagens estáticas adjacentes, e consegue atingir um alto índice de compressão. Existem diversas variantes do padrão. MPEG-1 foi projetado para se obter uma taxa de transmissão entre 1.2 a 2.0 Mbps, sendo adequado para uma boa qualidade visual (resolução média). MPEG-2 estabelece uma codificação com melhor qualidade de vídeo, cujas taxas alcançam entre 4 e 60 Mbps, para o suporte a HDTV. O MPEG-4 foi projetado com o objetivo de permitir taxas de transmissão extremamente baixas, que variam de 10 Kbps a 112 Kbps, mas com qualidade razoável.

Fatia de Tempo O conceito de fatia de tempo (*slot*) é utilizado geralmente em protocolos de difusão periódica para o auxílio em sua compreensão e formulação, e corresponde a um intervalo de tempo fixo arbitrado geralmente como o tempo de consumo do primeiro segmento.

Latência Tempo de espera, Latência de exibição, ou simplesmente latência são termos equivalentes. Podem ser definidos [18] como o tempo que o cliente aguarda para o início da exibição do vídeo desejado, compreendendo o intervalo entre o momento exato da requisição até o início da exibição propriamente dita.

Prefixo Corresponde a uma fração inicial de um vídeo (medida em unidades de tempo ou em número de quadros), cujo tamanho é definido de acordo com o esquema utilizado. Prefixos de vídeos populares podem ser armazenados previamente nas *set-top-boxes* dos usuários para eliminar o atraso decorrente dos protocolos de difusão e de difusão seletiva, técnica esta conhecida como *partial preload* ou pré-carregamento parcial. Os prefixos podem também ser armazenados em servidores intermediários, na técnica denominada *prefix caching*[45].

Segmentação Algumas abordagens de VoD, principalmente os protocolos de difusão periódica, utilizam uma divisão lógica dos vídeos em diversos segmentos contínuos representados por $S_1, S_2, \ldots, S_N$, cujos tamanhos, freqüências de exibição e alocação

nos diversos canais variam conforme a abordagem utilizada. A vantagem da segmentação na transmissão de um vídeo é que o tempo de espera consegue ser reduzido ao tempo de espera para a exibição (ou para recepção total) do primeiro segmento.

Taxa de Consumo Corresponde à taxa de processamento de dados da *set-top-box* para saída do vídeo[4]. Esta taxa de consumo b (Tabela 2.1) varia de acordo com o formato de vídeo utilizado, e é medida de referência para protocolos de difusão periódica, nos quais tanto a largura de banda requerida pelo cliente quanto pelo servidor são dadas em função de b. Por exemplo, ao se dizer que um determinado protocolo exige 5,92 canais, este valor na realidade indica que a soma da largura de banda alocada por todos os canais lógicos utilizados é equivalente a 5,92 vezes a taxa de consumo b.

Popularidade Costuma-se utilizar as nomenclaturas vídeos populares (*hot videos*) para os vídeos mais freqüentemente requisitados, e vídeos não populares (ou *cold videos*) para os restantes. A freqüência de acesso a filmes de entretenimento é uma das motivações para o compartilhamento de fluxo de banda passante, pois cerca de 80% da demanda responde pelos 10 a 20 filmes mais populares, como por exemplo os recém-lançados no mercado. Pode-se caracterizar esta demanda de filmes por uma distribuição de Zipf com um parâmetro de distribuição equivalente a 0.271 [8, 5]. Em uma distribuição de Zipf [52], ao classificar os vídeos de acordo com sua freqüência de requisições (vídeo um é o mais requisitado, vídeo número dois é o segundo mais requisitado, e assim sucessivamente). A freqüência f_i de solicitações para o vídeo i é dada por $f_i = \frac{c}{i^{(1-\theta)}}$, onde θ corresponde ao parâmetro da distribuição[1] e c equivale a uma constante de normalização.

2.1.6 Taxonomia

A principal classificação de esquemas para VoD é voltada à interatividade com o usuário final [31]. Estes esquemas estão dispostos conforme as seguintes classificações:

TVoD - True Video on Demand Modalidade de serviços considerada mais ideal, onde além da seleção do vídeo o usuário pode também executar quaisquer operações de VCR, como por exemplo avanço rápido, retrocesso, pausa e posicionamento aleatório. Neste esquema, cada usuário que requisita acesso a um vídeo terá um

[1] $0 \leq \theta \leq 1$, onde 0(zero) indica uma distribuição de Zipf pura, e 1(um) corresponde a uma distribuição uniforme [50]

canal lógico alocado exclusivamente[2]. Para permitir a interação total, nenhum outro usuário pode acessar este canal simultaneamente. A ocorrência de uma requisição sem a disponibilidade de canais lógicos implica em um descarte sumário do pedido, conhecido por *Blocking*. Uma variação deste esquema é a criação de uma fila com as requisições pendentes. Para isso, o usuário deve ser notificado do tempo de espera estimado, sendo que sua inclusão na fila só ocorre se o tempo máximo de espera for menor que o comunicado.

NVoD - Near Video on Demand Projetado de modo a oferecer um serviço de menor custo que o TVoD. Seu princípio básico consiste em iniciar um fluxo de cada vídeo popular a cada Δt minutos em canais distintos[3], de modo que um usuário que faz um pedido precisa esperar até Δt minutos para ser atendido (Ver seção 4.3, Protocolo de Difusão Balanceada).

PVoD - Partitioned Video on Demand Este esquema constitui-se da utilização conjunta dos anteriores, combinando suas vantagens. Parte dos canais é utilizada com o esquema NVoD para exibição dos vídeos populares, e os canais restantes são utilizados para exibição dos vídeos não populares através do esquema TVoD. Os usuários que requisitam vídeos exibidos pela partição NVoD são atendidos por ela, enquanto a partição TVoD atende os demais.

DAVoD - Dynamically Allocated Video on Demand Possui um comportamento similar ao PVoD, permitindo o deslocamento dinâmico de usuários, da partição NVoD para TVoD por ocasião da solicitação de operações de VCR.

No-VoD O usuário é passivo e não possui nenhum controle sobre a sessão, como em uma transmissão de TV convencional [30].

2.2 Técnicas de redução de banda passante

O oferecimento de serviços em VoD em larga escala depende em grande parte do controle de admissão e gerenciamento dos recursos alocáveis; porém, em virtude da grande demanda de banda passante requerida, a utilização de outras estratégias torna-se necessária.

[2]É possível que um usuário inicialmente seja vinculado a um canal de difusão seletiva. Entretanto, ao solicitar uma operação VCR, o usuário terá perdido a sincronização com o restante do grupo, requerendo para si um canal exclusivo.

[3]cada fluxo possui Δt minutos de atraso/avanço de exibição em relação ao fluxo anterior/posterior

Nos Estados Unidos (continental) estima-se que o número de aparelhos de TV ligados simlutaneamente em horários de pico totaliza 77 milhões. Num sistema de VoD com alocação de fluxos sob demanda, o montante de banda passante requerido chegaria, utilizando a taxa de pico, a 462 Tbps para a codificação MPEG-2 NTSC (6 Mbps/fluxo), 770 Tbps utilizando JPEG NTSC (10 Mbps/fluxo) e 1,54 Pbps para MPEG-2 HDTV (20 Mbps/fluxo). Estes altos requisitos evidenciam a necessidade da utilização de técnicas de redução da demanda de banda passante, seja isoladamente ou de forma conjunta[14].

Replicação, *caching* e *Prefix caching*

A técnica de replicação consiste na distribuição de diversas cópias de um mesmo objeto de vídeo em vários servidores, a fim de reduzir o custo de transmissão (volume de dados). Como o tamanho dos vídeos é substancialmente maior que os objetos atuais mais comuns na Internet (textos e imagens estáticas), a utilização de cache para armazenar todo o vídeo torna-se inviável. Por outro lado, se não houver um armazenamento prévio em cache, a latência será maior e será necessária uma largura de banda do servidor maior que a necessária. Uma solução para este problema foi apresentada em [48, 20] e baseia-se no armazenamento somente da parte inicial — i.e., o prefixo — do vídeo. Desta forma, torna-se possível o armazenamento de prefixos de uma variedade bem maior de vídeos ao mesmo tempo em que a latência de exibição dos mesmos é reduzida, pois enquanto a cache começa a servir os quadros iniciais do vídeo, os quadros restantes vão sendo requisitados ao servidor de vídeos. O tamanho do prefixo exerce uma influência sobre o compromisso entre o espaço do buffer de armazenamento dos vídeos e o requisito de latência máxima a ser imputada ao cliente.

Bridging

A técnica de *bridging* [19] consiste no armazenamento (em um *buffer* de memória) dos últimos k minutos de um vídeo, a fim de atender requisições — seja por novos fluxos ou pela retomada de exibição após operações VCR — num intervalo até k minutos após o quadro do filme em exibição.

2.2.1 Classificação de abordagens

As abordagens de transmissão, por seu foco principal, podem ser classificadas [17] em:

Voltada a usuários (*user-centered*) , onde os canais são alocados para usuários específicos, conforme suas requisições. Os protocolos baseados em difusão seletiva estão enquadrados nesta categoria. Estes protocolos também são denominados de natureza reativa[21], uma vez que estes protocolos transmitem os objetos de vídeo somente mediante requisições. Esta abordagem é ideal para os vídeos que não possuem uma taxa de requisição muito alta.

Voltada a dados (*data-centered*) , onde os algoritmos dedicam canais exclusivos à transmissão contínua de segmentos de vídeos. Por requerer largura de banda constante (o que garante escalabilidade), estes esquemas são chamados de proativos (*proactive*), e estão mais associados ao esquema NVoD e aos protocolos que utilizam difusão periódica.

2.2.2 Interatividade em VoD

Sistemas de Vídeo sob Demanda podem oferecer operações VCR, estas geralmente associadas ao TVoD, e podem ser obtidas através de canais suplementares dos servidores, alocadas especialmente para estes propósitos [15].

2.2.3 Notação

Para apresentar nas próximas seções os diferentes protocolos utilizados em VoD, será utilizada a seguinte notação:

2.3 Compartilhamento de Fluxo Baseado em Difusão Seletiva (Multicast)

A abordagem de difusão periódica não é muito eficiente para vídeos que não são solicitados com alta freqüência, pois continuamente transmitem um vídeo que pode não ter sido solicitado no momento. Abordagens reativas envolvem o uso de difusão seletiva, e tentam reduzir a utilização de largura de banda transmitindo os vídeos somente quando estes forem solicitados, e compartilhando as partes dos vídeos comuns a diferentes transmissões.

B	Largura de banda total do servidor
S	Duração total do vídeo
V	Número de vídeos da coleção disponíveis para escolha
w	latência de acesso [4]
f	Duração de tempo correspondente a uma fatia de tempo
b	Taxa de consumo do vídeo
C	Número de canais lógicos alocados para a transmissão de um vídeo
N	Número de segmentos do vídeo
S_i	Representa o i-ésimo segmento do vídeo
b_i	largura de banda lógica alocada para o canal lógico i

Tabela 2.1: Notação utilizada para descrever as abordagens utilizadas em VoD

2.3.1 Batching

Técnicas de Batching baseiam-se na retenção de requisições por um determinado período (denominado *Janela* ou *Intervalo*), para agrupá-las em um único fluxo de transmissão de difusão seletiva [11]. Há um compromisso, portanto, entre a latência e a maximização do número de usuários a serem servidos pela sessão. As seguintes políticas, propostas na literatura, estão associadas ao *Batching*:

Espera forçada Exige que pelo menos uma das requisições seja retida durante a janela de *Batching*. Ao final da mesma, aloca-se um fluxo para o vídeo solicitado. Este funcionamento pode causar, nos horários de pico, o fenômento da formação de ciclos, que consiste em momentos de intensa alocação de canais seguidos por longos intervalos de espera, isto é, grandes índices de rejeição de pedidos.

Controle de Taxa Puro Política que *i*) impõe um limite superior na taxa a qual canais podem ser alocados, e *ii*) limita o número total de canais que podem ser alocados durante um intervalo de tempo fixo (*intervalo de medida*) com o objetivo de evitar a formação de ciclos.

Controle de Taxa Desviado Difere da política anterior, pois permite que a taxa de alocação máxima com o objetivo de evitar o abandono de usuários (sem serviço) caso um canal não seja alocado em um curto intervalo de tempo.

Maior Tamanho de Fila - MQL (*Maximum Queue Length*) Busca maximizar o efeito de *Batching* alocando o fluxo para a fila cujo número de requisições pendentes seja maior.

Fila de Maior Índice - MFQL (*Maximum Factored Queue Length*) Define-se um índice que determina qual das filas de requisições receberá o fluxo a ser alocado, este índice representa um compromisso entre a maximização do efeito de *Batching* e o tempo de espera na fila. Dessa forma, busca-se maximizar a utilização dos recursos do sistema sem desprezar os vídeos menos populares.

Primeiro a Chegar, Primeiro a ser Atendido - FCFS (*First Come First Served*) Nesta política é considerado principalmente o aspecto de justiça entre os vídeos (populares e não populares) quando da alocação de fluxos. Assim sendo, todas as requisições são inseridas em uma única fila e a alocação é realizada conforme a ordem de chegada. Quando um canal torna-se disponível, a requisição da cabeça da fila é atendida e todas as outras requisições para o mesmo filme são servidas por este canal. Apesar de ser justa, como não leva em consideração o tamanho do grupo (*batch*), esta política possui um menor desempenho em relação a redução de banda passante.

Primeiro a Chegar, Primeiro a ser Atendido-n Extensão da política FCFS, onde canais dedicados são atribuídos aos n filmes mais populares, de tal forma que um novo fluxo para estes n filmes pode ser iniciado a cada w segundos. Assim o tempo de espera é no máximo w segundos para estes filmes. Além disso, como os filmes populares possuem taxa de requisições altas, cada fluxo serve mais de um usuário. Se o tempo w for ajustado apropriadamente à taxa de chegada, os recursos do sistema podem ser substancialemten reduzidos.

Paradigma de Tolerância de Espera O objetivo é maximizar o efeito do *Batching*, definindo um *intervalo de espera* (*wait threshold*), um intervalo de tempo no qual as requisições são retidas, explorando o tempo estimado de espera dos usuários. Definem-se duas classes de políticas, *Max_Batch* e *Min_Idle*: Em *Max_Batch*, um vídeo é escalonado para alocação se pelo menos uma de suas requisições foi retida durante o intervalo de espera. Caso mais de um vídeo satisfaça esta condição, aloca-se o fluxo utilizando-se o critério de maior tamanho de fila (*Max Batch MQL* - MBQ) ou de maior número estimado de abandonos (*Max Batch with Minimal Loss* - BML). Caso contrário, um fluxo permanece disponível até que uma requisição satisfaça a este critério. Em *Min_Idle*, divide-se os vídeos em dois conjuntos: populares, $\mathcal{H}$ (*hot videos*), e não populares, $\mathcal{C}$ (*cold videos*). O *Batching* é aplicado somente sobre o conjunto $\mathcal{H}$, sem restrição de tempo mínimo de espera no conjunto $\mathcal{C}$. Um vídeo é inserido em $\mathcal{H}$ se três requisitos forem atendidos: *i*) deve ser popular, *ii*) sua

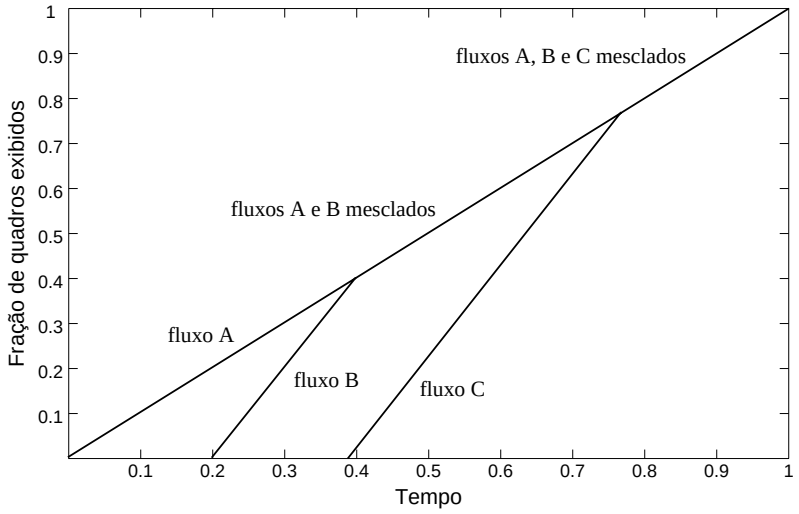

Figura 2.2: Exemplo de mesclagem de fluxos em *Piggybacking*

fila possuir mais de uma requisição, e *iii*) a única requisição de sua fila exceder o intervalo de espera. Se o conjunto $\mathcal{H}$ não estiver vazio, aloca-se um vídeo utilizando-se o critério de maior tamanho de fila (*Min Idle MQL* - IMQ) ou maior número esperado de perdas (*Min Idle with Minimal Loss* - IML). Caso contrário, aloca-se um fluxo para um vídeo do conjunto $\mathcal{C}$utilizando-se a política FCFS.

***Look-Ahead-Maximize-Batch* - LAMB** Política que admite o maior número de usuários dentre as políticas apresentadas, LAMB maximiza o efeito de *batching* utilizando conhecimento sobre o comportamento de abandono dos usuários.

2.3.2 Piggybacking

Piggybacking[7, 41, 19] baseia-se no fato de que usuários não percebem alterações de até 5% da taxa de exibição do vídeo, sendo as requisições prontamente atendidas pelo servidor. Quando um usuário requisita um vídeo que já esteja sendo transmitido, inicia-se a transmissão de um novo fluxo, de forma que este novo fluxo possua uma taxa de exibição maior que o fluxo existente. Em um dado momento (indicado na Figura 2.2), os fluxos são sincronizados, ou seja, os quadros de ambos fluxos serão exibidos no mesmo instante: é quando ocorre a mesclagem dos fluxos, onde um dos canais é liberado e o outro é compartilhado.

2.3.3 Integração de *Batching* e *Piggybacking*

A integração entre as técnicas de *Batching* e de *Piggybacking*[7] pode trazer um ganho de até 20% de performance em relação a uma técnica isolada. Porém, devem-se determinar políticas adequadas para a integração, pois algumas políticas requerem intervalos mínimos entre a alocação de fluxos (e.g. *Batch* MBQ), e *Piggybacking*, ao contrário, somente sobrepõem fluxos que distem um número máximo de quadros.

2.3.4 Patching

Um cliente que faça uma requisição de um vídeo cuja transmissão já esteja em andamento pode utilizar esta transmissão, armazenando os quadros desta transmissão no *buffer* de sua *Set-Top-Box* para exibí-los posteriormente. Desta forma, só é necessária a transmissão dos quadros iniciais, servidas em um fluxo exclusivo, o que garante início imediato de exibição. Diferentes esquemas de *patching* decidem até que ponto é mais vantajoso utilizar uma transmissão existente ou iniciar um novo fluxo.

Em patching[23, 3, 2], o servidor envia todo o fluxo do vídeo sequencialmente para o primeiro cliente que o solicitou. Novas requisições de clientes aproveitam esta transmissão já existente em um canal de difusão seletiva para exibição futura (o que torna necessário que os clientes possuam algum dispositivo de armazenamento temporário para armazenar estes quadros) e recebem individualmente apenas os quadros iniciais faltantes, diretamente do servidor.

2.3.5 PBR - Patching Buffer Reuse

Os algoritmos tradicionais de patching limitam o cliente a uma única sessão existente por vídeo, o que obriga os clientes a possuírem uma quantidade excessiva de buffer. O PBR patching[43] pode decidir quando aproveitar uma sessão existente ou iniciar uma nova sessão, isto é, avalia o compromisso entre maior uso de largura de banda do servidor ou garantir um tamanho de cache menor para os clientes.

2.3.6 Catching

Catching[18] é uma técnica voltada para a exibição de *hot videos*(vídeos freqüentemente acessados). Um certo número de canais dedicados para *multicast* é alocado para exibição periódica do vídeos via difusão. Um cliente que queira assistir imediatamente o vídeo une-se à transmissão *broadcast* e vai armazenando o vídeo em sua cache local. Ao mesmo

tempo, o cliente solicita ao servidor uma requisição para obter os quadros iniciais faltantes do vídeo. Estes quadros são servidos através de uma conexão unicast, e quando recebidos pelo cliente podem então ser o vídeo pode ser exibidos. O número de canais alocado para difusão depende da técnica a ser empregada. Por exemplo, em [18] utiliza-se o GDB.

2.3.7 Multicast com Caching (Mcache)

A idéia central do Mcache[39] é integrar *batching*, *patching* e técnicas de *prefix caching* — tendo como base a utilização de difusão seletiva — aproveitando as vantagens e minimizando as desvantagens de cada técnica. O cliente envia a requisição de um vídeo para o seu *proxy*. Este, ao receber a requisição, repassa ao servidor a requisição do vídeo desejado e simultaneamente atende o cliente, fornecendo-lhe o prefixo, diminuindo assim a latência de exibição. Enquanto isso, o servidor vai acumulando (*batching*) as requisições que chegam, pelo tempo de duração da exibição do prefixo contido na cache. Após a exibição do prefixo, todas as requisições reunidas pelo *batching* são servidas por apenas um canal de difusão seletiva, proporcionando uma melhor utilização da largura de banda. Requisições de clientes que chegarem no momento da recepção de um segmento S_i posterior ao início da sessão de difusão seletiva ainda podem unir-se a esta mesma sessão via *patching*: neste caso cada cliente recebe a parte inicial da cache, e novamente o servidor do vídeo faz o *batching* das requisições de *patching*, transmitindo todos os segmentos faltantes (desde S_2 até S_i, uma vez que o segmento S_1 já foi provido pela cache) em um canal de difusão seletiva.

2.4 Compartilhamento de Fluxo Baseado em Difusão

Algoritmos baseados em difusão periódica têm sua largura banda dedicada aos objetos de vídeo e não aos usuários[1, 2]. A conseqüência é que esta abordagem torna-se atraente na medida em que aumenta-se o número de clientes e o número de requisições para exibição do mesmo objeto de vídeo, garantindo maior escalabilidade. Dados de uma porção posterior do vídeo podem ser recebidos pelo cliente, armazenando quadros futuros em suas caches[44].

2.5 Síntese do Capítulo

Um sistema de Vídeo sob Demanda permite a seleção de vídeos dentre vastas coleções para exibição, através de uma rede de computadores. Em uma arquitetura de um sistema VoD, destacam-se os seguintes componentes: o servidor de vídeo, a rede de de distribuição, a rede de acesso local e o equipamento do usuário (*set-top-box*).

Objetos de vídeo requerem uma enorme demanda de largura de banda. Para reduzir esta demanda, foram propostas diversas técnicas, onde as principais são baseadas no compartilhamento de fluxo através de transmissão via difusão ou difusão seletiva.

Protocolos baseados em difusão seletiva possuem uma característica reativa e são ideais para vídeos com taxas de requisição baixas ou médias. Podem-se destacar o *Batching*, *Piggybacking*, *Patching* e *Catching*.

Outra classe de protocolos é baseada na divisão de um vídeo e na difusão periódica de seus segmentos. O usuário participa de transmissões em andamento, conectando-se a um ou mais canais e armazenando segmentos futuros em um *buffer* para posterior exibição. Esta abordagem cria a necessidade de se esperar pela exibição (NVoD), mas em compensação possui maior escalabilidade.

Capítulo 3

Uma Introdução aos Algoritmos Genéticos

Neste capítulo uma breve introdução aos Algoritmos Genéticos (AG) é apresentada, dado que esta técnica de otimização será utilizada para a determinação de configurações ótimas em protocolos de difusãoperiódica. Algoritmos genéticos correspondem a uma heurística utilizada para obtenção de resultados aproximados em problemas que requerem grande espaço de busca, como os estudados no presente trabalho.

A Seção 3.1 introduz o tema e a analogia entre sistemas evolutivos naturais e computacionais. A Seção 3.2 mostra o funcionamento geral de um AG. As seções seguintes detalham as etapas presentes na elaboração de um AG: a estruturação do problema (Seção 3.3), geração de uma população inicial (Seção 3.4), definição de mecanismos de avaliação dos cromossomos (Seção 3.5), escolha dos operadores genéticos (Seção 3.6) e, finalmente, definição dos parâmetros genéticos (Seção 3.7).

3.1 Introdução

Algoritmos genéticos [9, 27, 13, 28, 49], ou AG, correspondem a uma área dos algoritmos evolucionários, os quais procuram encontrar soluções aproximadas para problemas de grande complexidade computacional através de um processo de evolução simulada análogo ao processo de seleção natural, na qual indivíduos reproduzem-se (trocando informações genéticas), sofrem mutações, e os genes dos indivíduos mais adaptados possuem maior chance de sobrevivência através das gerações.

Além da obtenção de soluções aproximadas em problemas de otimização, diversas

outras aplicações que podem ser beneficiadas com a utilização de algoritmos genéticos são encontradas na literatura. Dentre elas, incluem-se o aprendizado de máquinas (*machine learning*), modelagem de fenômenos ecológicos e aplicações em aviação.

3.2 Estrutura geral de um AG

Os algoritmos genéticos utilizam uma abordagem diferenciada no que tange ao espaço de busca. Eles utilizam uma população de indivíduos, onde cada indivíduo represeta um ponto dentro do espaço de busca de um determinado problema. Cada indivíduo possui um grau de aptidão, função que indica maior probabilidade deste ser selecionado, tanto para a geração seguinte quanto para a utilização de seus genes para a formação de novos indivíduos.

Todo algoritmo genético deve possuir as seguintes etapas: representação de uma solução em uma estrutura de cromossomo, mecanismo para avaliação dos cromossomos (segundo suas aptidões), definição dos parâmetros genéticos, geração de uma população inicial de cromossomos e mecanismos para a geração, reprodução e/ou alteração de composição dos cromossomos.

Concluídas as etapas preliminares (detalhadas nas seções seguintes), entra a fase de execução do algoritmo, como o indicado na Figura 3.1, onde o número da geração é representado pela variável `t`. Em primeiro lugar, é gerada uma população inicial, sendo gerada também uma avaliação para cada indivíduo. A seguir, são realizadas operações genéticas (textttmutação(t), `cruzamento(t)`) nos indivíduos, gerando assim uma nova população. Posteriormente, são selecionados os indivíduos que farão parte da próxima geração (`seleção(t)`). O ciclo recomeça pelas operações genéticas, até que a `CONDIÇÃO_DE_PARADA` seja satisfeita. Esta é geralmente expressa em função do número de gerações, devendo este número ser suficientemente grande para que a solução possa convergir para um valor ótimo.

3.3 Representação de uma solução em uma estrutura de cromossomo

Nesta etapa é realizado o mapeamento de uma solução do problema a ser resolvido em uma estrutura que possibilite a utilização de técnicas de AG, denominada cromossomo. Assume-se que cada indivíduo possui um cromossomo, dividido em vários genes. Estes

```
procedimento AG;

início
   t := 0
   inicia população(t);

   repita

      mutação(t);
      cruzamento(t);

      seleção(t);
      t := t+1;

   até que CONDIÇÃO_DE_PARADA

fim.
```

Figura 3.1: Pseudo-código de um Algoritmo Genético

genes devem conter uma representação de uma solução do problema a ser otimizado. Exemplificando, pode-se associar um gene para cada variável de decisão em um problema de otimização. Os genes devem possibilitar a obtenção de um valor de aptidão (Seção 3.5) que indicará a qualidade da solução.

Na sua forma original, os algoritmos genéticos tradicionais utilizam representações binárias para fazer a associação de uma solução do problema a ser abordado. Em aplicações com várias restrições, a representação binária pode não ser a mais apropriada. Pode-se citar como um exemplo a representação de uma solução no protocolo PDPH (Seção 5.5.5). Considere a geração de um novo indivíduo através de mutação de um cromossomo visto na Figura 3.3, em um gene representado pelo parâmetro m_1. Considere também que o valor deste parâmetro no qual o valor máximo permitido seja 100. Codificado de forma binária, uma mutação de algum bit deste gene poderia levar a um valor fora do escopo original (por exemplo, 127) que simplesmente não atenda à solução requisitada. Para este tipo de problema, a melhor alternativa é codificar cada gene como um parâmetro da solução, representado por um número inteiro ou outra que melhor convier.

3.4 Geração de uma população inicial de cromossomos

Corresponde à fase de geração dos indivíduos para compor uma população inicial. Para cada indivíduo, os genes são escolhidos de forma aleatória entre seus possíveis valores. Se o cromossomo gerado não satisfizer todas as restrições, seu indivíduo é descartado e o processo reinicia, até que todos os indivíduos sejam formados. A geração aleatória garante uma população diversificada, fato importante para a eficiência dos cruzamentos genéticos.

3.5 Mecanismo para avaliação dos cromossomos, segundo suas aptidões

O problema a ser solucionado deve ser definido e capturado em uma função objetivo que indique a aptidão de qualquer solução em potencial. Esta função é extremamente importante para um bom desempenho do AG, pois trata-se da ligação do algoritmo com o problema. *Fitness*, ou Função de aptidão, é um valor associado a cada indivíduo e que demonstra sua capacidade de sobrevivência. Quanto mais próximo do valor ótimo estiver a solução representada por um indivíduo, maior será o valor de sua aptidão e, conseqüentemente, maiores serão as chances de ser selecionado para a próxima geração.

Em problemas de otimização, a função de aptidão está diretamente relacionada com a função objetivo. Em alguns casos, recomenda-se a utilização de ajustes na função de aptidão de forma que esta permita que pequenas diferenças entre duas soluções possam ser ampliadas, aumentando as chances de seleção para a melhor solução, beneficiando-se principalmente quando a população atinge um grau maior de homogeneidade.

3.6 Operadores Genéticos: mecanismos de reprodução de cromossomos

Operadores genéticos possuem importância vital em AGs. Eles garantem a variabilidade das soluções no espaço de busca.

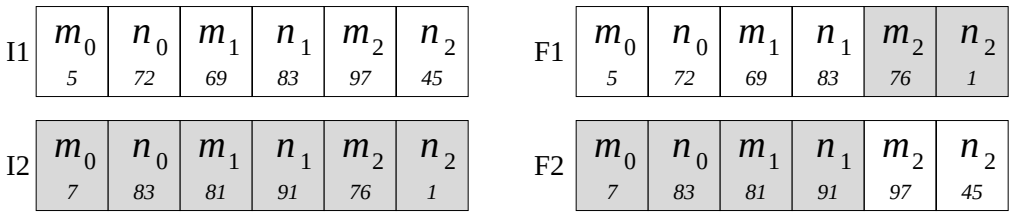

Figura 3.2: Operadores Genéticos: Cruzamento

Figura 3.3: Operadores Genéticos: Mutação

3.6.1 Cruzamento

Cruzamento (*crossover*) é uma operação genética que consiste na divisão do material genético de dois indivíduos diferentes, correspondendo à reprodução sexuada na natureza. Em AGs, os indivíduos são selecionados para cruzamento utilizando-se métodos como o da roleta, e seu funcionamento é indicado na Figura 3.2. Um ponto de corte (Na figura, correspondente ao quarto gene, n_1) é selecionado aleatoriamente, e o primeiro filho é gerado com genes de um dos pais até o ponto de corte, recebendo os genes finais do outro pai. Com o outro filho o raciocínio é inverso. Esta operação é tida como a principal operação genética, análoga à reprodução sexuada encontrada na natureza: indivíduos combinam seus códigos genéticos, criando uma variedade nas soluções.

Uma variante desta operação genética é o chamado cruzamento em dois pontos (*two-point crossover*), onde são eleitos dois pontos para cruzamento e somente os genes incluídos na faixa entre os dois pontos de corte são trocados.

3.6.2 Mutação

Um novo indivíduo é gerado por mutação a partir da modificação de um ou mais de seus genes (Figura 3.3). Para cada gene, é verificada a possibilidade de sua mutação do mesmo. Em caso positivo, é gerado um novo valor para o gene e, satisfazendo-se as restrições do problema, um novo indivíduo é gerado.

3.6.3 Seleção de indivíduos

A seleção em si não é uma operação genética, mas é a base para uma das operações mais importantes: a reprodução, que consiste na seleção dos indivíduos para a próxima geração.

A técnica da roleta para a seleção dos pais (*Roullete Whell Parent Selection*) é a mais utilizada entre os métodos de seleção, quando a função de aptidão for sempre positiva. Esta técnica consiste em três passos:

1. Os graus de aptidão são somados e acumulados;

A cada indivíduo que se queira selecionar, utiliza-se os seguintes passos:

2. Um número aleatório é gerado entre zero e o maior ajuste acumulado; e, finalmente,

3. é selecionado o cromossomo que possuir o menor grau de aptidão acumulado que seja maior ou igual ao número aleatório gerado.

A Tabela 3.1 exemplifica o processo de seleção via roleta. A tabela superior corresponde ao primeiro passo da seleção, onde cada cromossomo possui seu próprio valor de aptidão, que corresponderá à sua própria probabilidade de sobrevivência em relação aos demais. No exemplo, a cada sorteio o cromossomo de número cinco (que possui a maior aptidão dentre os demais) tem chance de 5 em 69 (7,2%) de ser selecionado a cada sorteio. Um mesmo cromossomo pode ser selecionado várias vezes.

Cromossomo	1	2	3	4	5	6	7	8	9	10
Aptidão	8	3	2	5	6	10	8	6	14	3
Aptidão Acumulada	8	11	13	18	24	34	46	52	66	69

Número aleatório	14	11	45	12	49	49	69	1	25	9
Cromossomo escolhido	4	2	7	3	8	8	10	1	6	2

Tabela 3.1: Exemplo da técnica da roleta para seleção

3.6.4 Reprodução Seletiva ou Elitismo

Na seleção natural, não há garantias que o indivíduo mais apto garanta sua sobrevivência automaticamente. No exemplo anterior (Tabela 3.1), o indivíduo de número cinco não seria selecionado para a próxima geração, embora possuísse o maior grau de aptidão dentre os demais. A reprodução seletiva é uma técnica opcional (e amplamente utilizada) onde o indivíduo com o valor de aptidão mais alto é copiado para a geração seguinte. Trata-se

de uma forma de não perder soluções potenciais em estágios intermediários do algoritmo, melhorando o desempenho do algoritmo genético.

3.6.5 Outros operadores genéticos

Outros operadores genéticos são definidos para o paradigma de programação genética. Entre eles, citam-se Inversão, Permutação, Edição, Encapsulamento, e Dizimação, detalhados na referência [27].

3.7 Parâmetros Genéticos

Os parâmetros utilizados em algoritmos genéticos influenciam na qualidade e no tempo de execução dos algoritmos genéticos. Cada domínio de utilização requer parâmetros genéticos diferentes. Os principais parâmetros genéticos são listados a seguir:

- **Tamanho da População** - Corresponde ao grau de cobertura do espaço de busca. Um aumento no número de indivíduos eleva este grau de cobertura, mas em compensação traz um acréscimo no tempo de execução para processamento de cada geração e, conseqüentemente, no tempo final de processamento.

- **Número de Gerações** - Equivale ao número de iterações (compreendendo as operações genéticas e seleção de indivíduos para nova geração) executadas pelo algoritmo. O valor deve ser grande o suficiente para as soluções possam convergir e depende da aplicação utilizada.

- **Taxa de Cruzamento** - Por se tratar da principal operação genética, os valores comuns para este parâmetro geralmente são altos, geralmente acima de 60% e chegando até 90%.

- **Taxa de Mutação** - Indica qual a fração dos indivíduos sofrerá mutação em um de seus genes em uma geração.

  Valores extremamente baixos limitam o espaço de busca, conseqüentemente encontrando somente máximos locais e dificultando a obtenção de soluções melhores. Por outro lado, o aumento dos valores para este parâmetro tornam o algoritmo progressivamente aleatório, prejudicando sua performance. Num exemplo extremo, uma taxa de mutação de 100% transforma um algoritmo genético em uma busca aleatória, cujo

desempenho é totalmente insatisfatório. Taxas típicas de mutação geralmente são baixas, girando em torno de 2%.

- **Outros Parâmetros Qualitativos** Também são considerados parâmetros qualitativos em AGs [27] os próprios métodos utilizados tanto para seleção básica de indivíduos como a seleção para acasalamento (por exemplo, método diretamente proporcional à aptidão), a existência ou não de um ajuste para a função de aptidão, e a utilização ou não de estratégia elitista.

3.8 Síntese do Capítulo

Algoritmos genéticos fazem parte de um ramo de algoritmos evolucionários, e surgem como uma heurística para a obtenção de soluções aproximadas, especialmente em problemas de otimização e aprendizado de máquinas. A idéia principal baseia-se na simulação da reprodução das espécies naturais, onde há uma população de indivíduos, que se reproduzem (trocando material genético entre os indivíduos) durante várias gerações, de forma que os portadores de genes que possuírem maior capacidade de adaptação têm maiores chances de transmitir seus códigos genéticos para as gerações seguintes. Para possibilitar esta analogia, é necessário haver um mecanismo de avaliação dos cromossomos para indicar suas capacidades de sobrevivência. No contexto dos AGs, utiliza-se uma função de aptidão, que associa um valor ao cromossomo (e indica a qualidade da solução).

As principais operações genéticas utilizadas pelos algoritmos genéticos são o cruzamento e a mutação. A seleção de indivíduos (tanto para sobrevivência à próxima geração quanto para obtenção de pais para as operações) é feita de forma proporcional à aptidão de cada indivíduo.

Os AGs possuem parâmetros genéticos que determinam o tempo de resposta e a qualidade da solução, sendo estes parâmetro específicos para cada domínio.

Capítulo 4

Protocolos de Difusão Periódica

Em protocolos de difusão periódica, o servidor divide um vídeo em vários segmentos, e transmite-os em um conjunto de canais dedicados à transmissão destes segmentos; cada cliente recebe dados de vários canais ao mesmo tempo para armazenar/exibir os quadros posteriores. Todos os protocolos que utilizam difusão periódica, excetuando-se o Protocolo de Difusão Balanceada, possuem uma organização similar: dividem cada vídeo em N segmentos, transmitidos simultaneamente via difusão em diferentes canais lógicos (Figura 4.1).

O primeiro destes canais transmite apenas o primeiro segmento, de forma que se possa reduzir a latência de exibição do vídeo. Os primeiros protocolos desenvolvidos esperam o início de uma transmissão do primeiro segmento para então recebê-lo enquanto o exibe. Protocolos mais recentes recebem os segmentos no instante em que se unem à transmissão em andamento, requerendo então que cada segmento seja recuperado em sua totalidade antes de iniciar sua exibição. Em ambos casos, quanto menor for o tamanho deste primeiro segmento, menor será a latência de exibição.

Os outros canais transmitem os segmentos restantes com suas larguras de banda designadas. Quando usuários desejam assistir um vídeo, esperam pelo início do primeiro segmento do primeiro canal. Enquanto eles começam a assistir este segmento, suas *set top boxes* gravam dados de outros canais em cache para exibição posterior.

A Seção 4.1 apresenta a correspondência entre os canais lógicos utilizados pelos protocolos de difusão periódica e os canais físicos efetivamente transmitidos. A seguir, na Seção 4.2, é apresentada uma ferramenta para visualização e criação de protocolos, o mapa difusão largura de banda *vs.* tempo.

As Seções 4.3 a 4.15 referenciam os diversos protocolos de difusão periódica disponíveis na literatura de VoD.

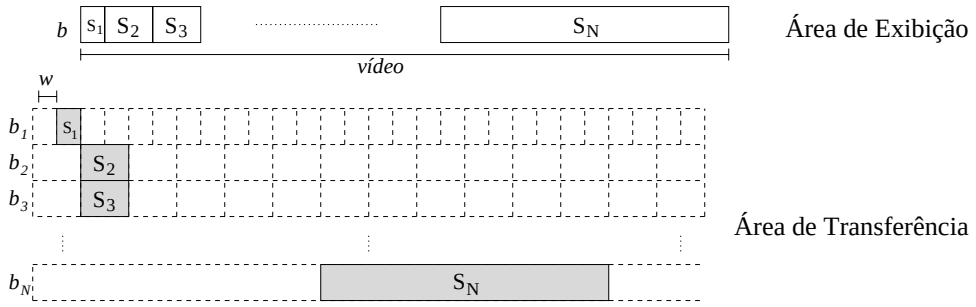

Figura 4.1: Esquema de um protocolo de difusão periódica

Na Seção 4.16, é abordado o pré-carregamento parcial de segmentos, uma técnica que utilizada em conjunto com algum protocolo de difusão periódica permite a exibição imediata do vídeo.

Por fim, a Seção 4.17 traz a definição de protocolo otimamente estruturado, bem como as regras que um protocolo deve seguir para sê-lo.

4.1 Multiplexação de canais

Protocolos de difusão periódica utilizam diversos canais para transmitir os segmentos. Um canal pode ser definido [24] como uma divisão lógica da largura de banda do servidor de vídeosvídeo.

A Figura 4.2 ilustra esta partição do canal físico (com largura de banda B) em C canais lógicos de largura de banda iguais a $\frac{B}{C}$, através de multiplexação de tempo [47]. Os dados representados por uma unidade de tempo passam a ser representados por C unidades de tempo. Desta forma, os dados da i-ésima fatia de tempo do canal físico são mapeados nas fatias de tempo de $\lfloor \frac{i}{K} \rfloor * K$ até $\lfloor \frac{i}{K} \rfloor * K + K$ do (i mod K)-ésimo canal lógico. O mapeamento contrário pode ser obtido através de sua função inversa: dados das unidades de tempo variando de $i * K$ até $(i+1) * K$ do canal lógico de número j pode ser mapeado para a $(i * K + j)$-ésima unidade de tempo do canal físico.

4.2 Mapa difusão largura de banda *versus* tempo

Os protocolos que utilizam difusão periódica podem ter seu funcionamento representado visualmente por uma abstração denominada mapa largura de banda-temporal [21] (ou simplesmente mapa de difusão). Neste mapa, os eixos x e y representam, respectivamente,

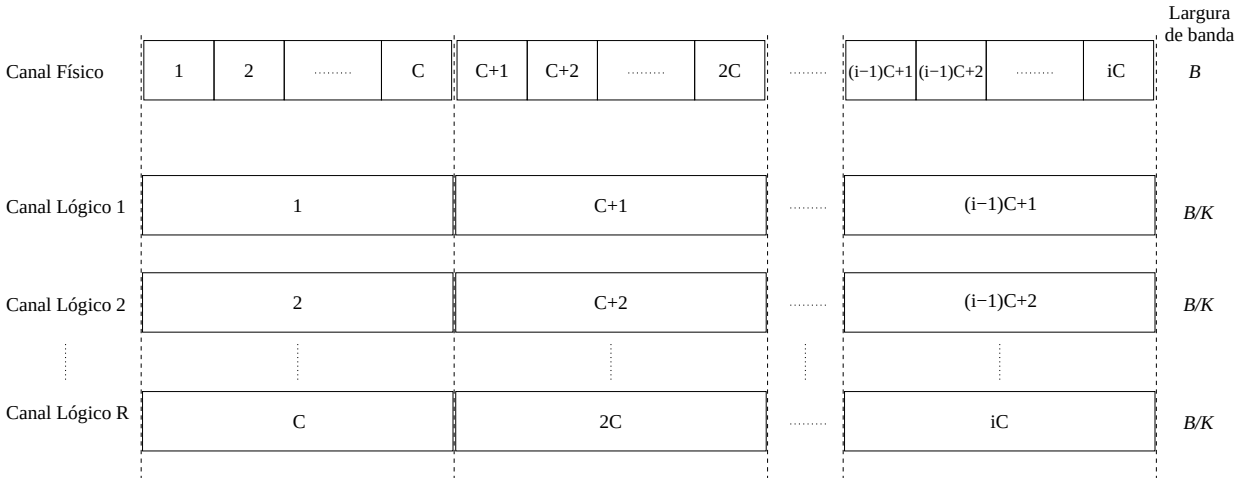

Figura 4.2: Divisão de um canal físico em canais lógicos

o tempo e a largura de banda utilizada pelo servidor. O mapa é composto por duas áreas principais (Figura 4.1).

A parte superior do mapa, denominado área de exibição (*playout area*) indica o momento da exibição de cada segmento pela *set-top-box* do usuário.

A parte inferior do mapa, denominada área de transferência (*broadcasting area*), representa os canais lógicos transmitidos pelo servidor durante o decorrer do tempo. As barras horizontais sobrepostas correspondem aos canais lógicos alocados pelo servidor e alturas indicam a largura de banda alocada a cada canal. As barras com contorno sólido e sombreadas indicam o período de recepção do segmento realizada por um cliente; as linhas verticais tracejadas, por sua vez, marcam os instantes de de ínicio e término da transmissão de um segmento.

4.3 Difusão Balanceada (*Staggered Broadcast*)

O Protocolo de Difusão Balanceada, (*Staggered Broadcast*), foi um dos primeiros propostos. É também um dos protocolos mais simples dentre os que utilizam difusão periódica. Neste protocolo, não há segmentação e cada um dos C canais alocados transmite todo o vídeo de forma contínua. Em seguida ao término da transmissão do vídeo, reinicia-se sua transmissão. Cada um dos canais transmite com a mesma largura de banda que a taxa de consumo do vídeo (b), mas com um atraso de $\frac{S}{C}$ unidades de tempo em relação ao canal anterior.

Os baixos requisitos de armazenamento (ou até mesmo nulos, em casos de vídeos CBR

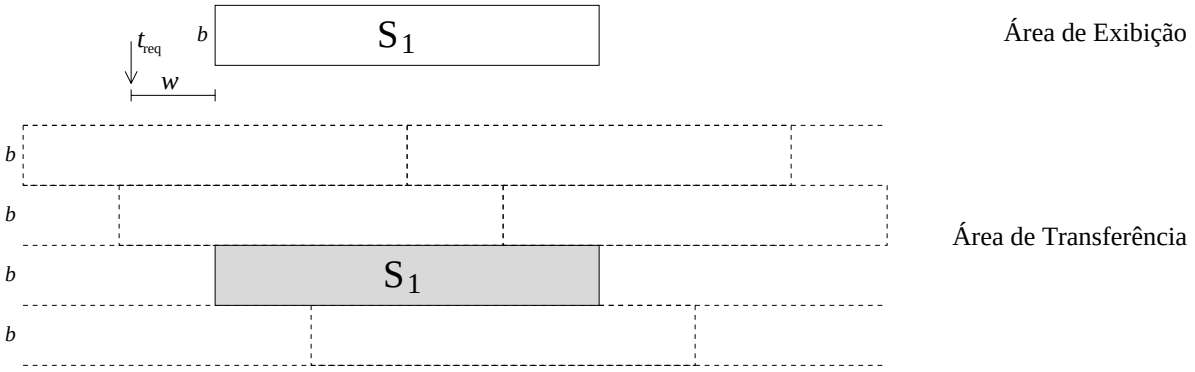

Figura 4.3: Mapa largura de banda *versus* tempo do Protocolo de Difusão Balanceada

transmitidos em redes que disponham de QoS) e o fato dos clientes só precisarem receber dados de um único canal tornam o Protocolo de Difusão Balanceada (PDB) muito simples e de fácil implementação, sendo utilizado atualmente em canais de TV por assinatura. Entretanto,o PDB peca pelo alto tempo de espera exigido e pela crescente ineficiência da utilização da largura de banda do servidor ao se tentar reduzir este tempo de espera. O tempo de espera é inversamente proporcional à largura de banda requerida pelo servidor. Na Figura 4.3, vê-se um exemplo do Protocolo de Difusão Balanceada. Para um filme de duas horas, com quatro canais alocados para sua transmissão, é exigido um tempo de espera máximo de trinta minutos e um tempo médio de quinze minutos. Neste exemplo, para se reduzir o tempo de espera máximo a dez minutos (um terço do tempo de espera anterior) é necessário triplicar o número de canais, requerendo então doze canais.

4.4 Protocolo de Difusão Piramidal (*Pyramid Broadcast*)

O Protocolo de Difusão Piramidal (*Pyramid Broadcasting*) [47] trouxe uma inovação dentre os protocolos de difusão periódica. A idéia central deste protocolo baseia-se na segmentação do vídeo para possibilitar a redução na latência de exibição, sendo adotada por todos os protocolos posteriores. Estes segmentos são alocados em canais lógicos, em uma relação um-para-um, e transmitidos via difusão de forma ininterrupta, isto é, ao término de cada transmissão de um segmento em um canal qualquer, reinicia-se a transmissão do mesmo segmento no mesmo canal. Cada cliente salva o próximo segmento de dados enquanto exibe o segmento atual (Figura 4.4). Através destas inovações, o Protocolo de Difusão Piramidal (PDP) inaugurou a primeira família de protocolos de difusão periódica,

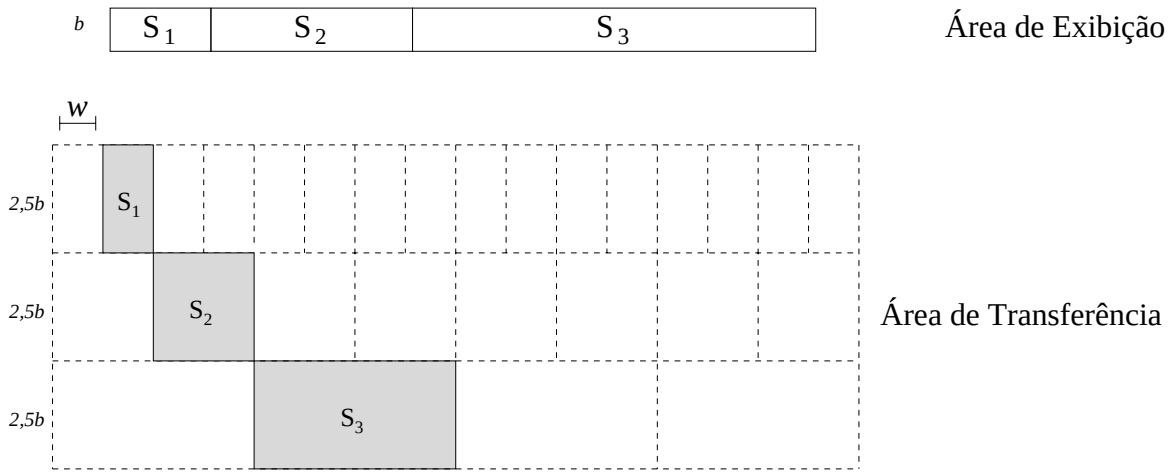

Figura 4.4: Mapa de difusão do Protocolo de Difusão Piramidal para $\alpha = 2.5$

onde o tamanho dos segmentos são crescentes e a largura de banda é a mesma para todos os canais.

O tamanho dos segmentos segue uma progressão geométrica com fator α , ou seja, [1, α, α^2, α^3, ...]. Cada segmento S_i, portanto, equivale a αS_{i-1} para $i \geq 2$. O primeiro segmento, por sua vez, é obtido pela expressão

$$S_1 = \frac{S(\alpha - 1)}{(\alpha^n - 1)}, \tag{4.1}$$

onde o valor ótimo de α para que o tempo de acesso seja minimizado é igual a 2,5.

A largura de banda é a mesma para todos os canais e é equivalente a α vezes a taxa de consumo b. Como a taxa de recepção é maior que a taxa de consumo, o tempo máximo de espera é reduzido a $\frac{S_1}{\alpha}$.

Apesar de ocasionar a redução do tempo de espera, esta alta taxa de transmissão por canal traz como conseqüência altos requisitos de armazenamento para o cliente — da ordem de 70% do vídeo — além do número de operações de I/O e da própria largura de banda total também atingirem níveis elevados, trazendo uma desvantagem em relação a outros protocolos desenvolvidos posteriormente.

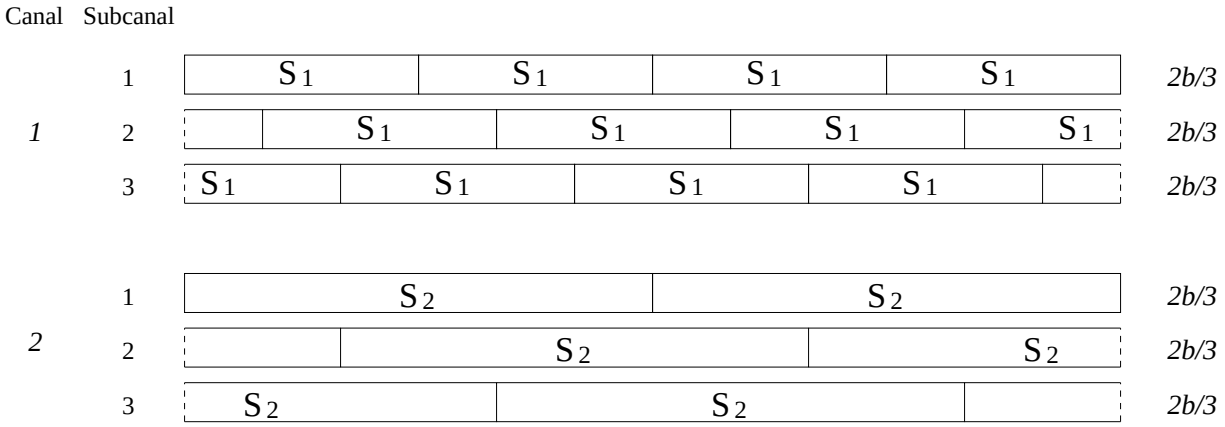

Figura 4.5: Exemplo de multiplexação dos canais no PDPBP

4.5 Protocolo de Difusão Piramidal Baseado em Permutações

O Protocolo de Difusão Piramidal Baseado em Permutações (*Permutation-based Pyramid Broadcasting*)[1] (PDPBP) tenta reduzir a alta demanda de armazenamento e largura de banda impostas ao cliente pelo seu antecessor, o Protocolo de Difusão Piramidal. A principal mudança neste protocolo é que, ao invés de se transmitir um segmento em um canal com grande largura de banda, multiplexa-se cada canal em P subcanais e os segmentos são transmitidos a uma taxa P vezes menor. Estes P subcanais transmitem o mesmo conteúdo que um segmento comum do DP, mas com o início dos segmentos atrasados em relação ao subcanal anterior, da mesma forma que no protocolo de Difusão Balanceada, para satisfazer as mesmas restrições de tempo que o PDP. A Figura 4.5 ilustra esta multiplexação, com parâmetros $\alpha = 2$ e $P = 3$ em um vídeo com somente dois segmentos.

4.6 Protocolo de Difusão Arranha-céu (*Skyscraper Broadcasting*)

O principal objetivo do Protocolo de Difusão Arranha-céu (*Skyscraper Broadcast*)[24] é atender a entrega de vídeos de forma que o cliente não seja obrigado a possuir altos requisitos de largura de banda e de armazenamento: o protocolo nunca utiliza mais do que dois canais simultâneos e requer um espaço de armazenamento temporário entre 3%

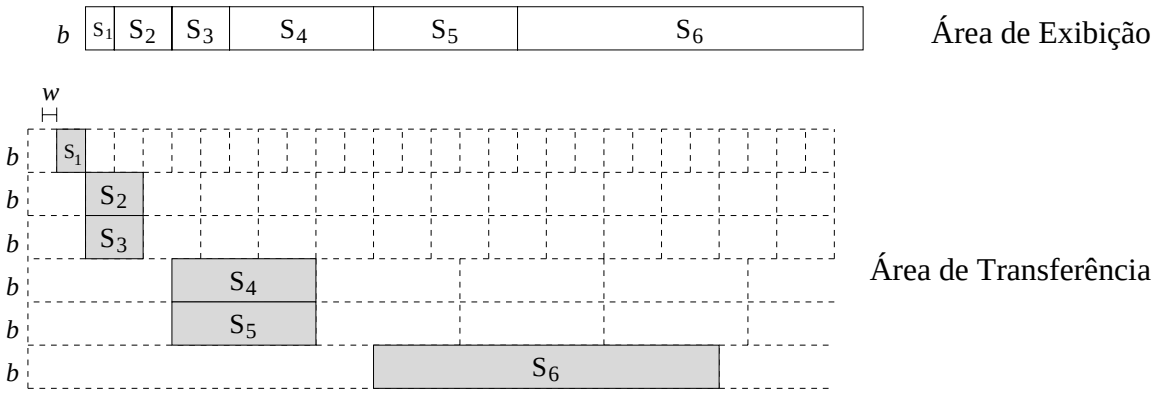

Figura 4.6: Mapa do Protocolo de Difusão Arranha-céu

e 30% do vídeo. O Protocolo de Difusão Arranha-céu (PDA) segue os mesmos princípios que o PDP, onde cada canal transmite um único segmento utilizando sempre a mesma largura de banda. O PDA também utiliza uma segmentação de tamanho crescente, mas de acordo com a seguinte função:

$$f(n) = \begin{cases} 1 & n = 1; \\ 2 & n = 2, 3; \\ 2f(n-1)+1 & n \bmod 4 = 0; \\ f(n-1) & n \bmod 4 = 1; \\ 2f(n-1)+2 & n \bmod 4 = 2; \\ f(n-1) & n \bmod 4 = 3. \end{cases}$$

Exemplificando, os tamanhos dos segmentos seguem a série:

$$[1, 2, 2, 5, 5, 12, 12, 25, 25, 52, 52, \ldots]$$

Neste protocolo, é incluído também um parâmetro W, correspondendo ao tamanho máximo possível para um segmento. Cada segmento S_i, portanto, equivale a:

$$S_i = \min\{f(i)S_1, W\} \tag{4.2}$$

A função de crescimento dos segmentos foi idealizada para poder atender sempre ao requisito de largura de banda. Na Figura 4.6, pode-se notar que os canais são sempre recebidos aos pares, um canal de número ímpar e outro de número par, à exceção do primeiro canal, que é recuperado isoladamente. Diferentemente dos protocolos anteriores, a transmissão de cada canal é efetuada a mesma taxa e de consumo.

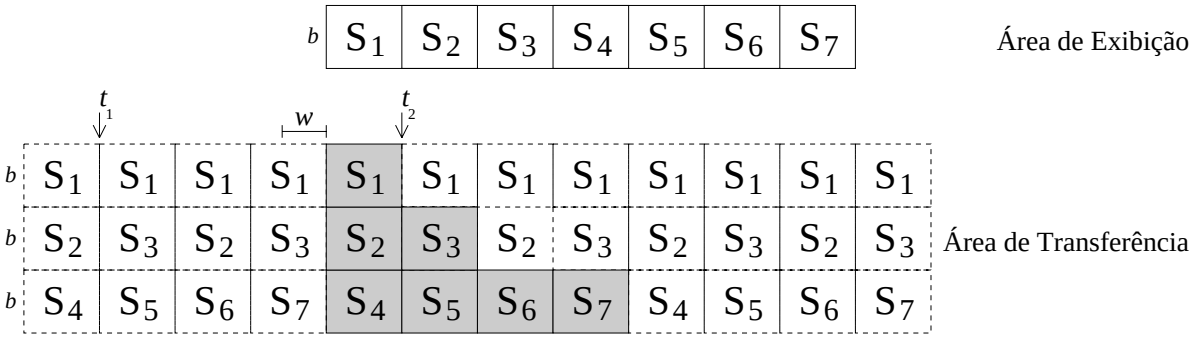

Figura 4.7: Mapa do Protocolo de Difusão Rápida

4.7 Greedy Disk-conserving Broadcast (GDB)

GDB [16] faz parte de uma categoria de esquemas chamada de *disk-conserving broadcasting schemes*, cuja idéia principal é conservar o espaço em disco do cliente de forma que este receba dados de forma mais tarde possível.

4.8 Protocolo de Difusão Rápida

O protocolo de Difusão Rápida (*Fast Broadcasting*) [26] utiliza $2^C - 1$ segmentos de tamanho igual, e cada canal i transmite periodicamente os segmentos $S_{2^{i-1}} \ldots S_{2^i}$ em seqüência, como indicado na Figura 4.7.

Como a transmissão dos segmentos é feita de forma contígua dentro de um mesmo canal, costuma-se também apresentar este protocolo [21] como um protocolo da família piramidal, sendo cada i-ésimo canal transmite 2^i segmentos. por sua vez, divide o vídeo em segmentos cujos tamanhos seguem a série:

$$[1, 2, 2^2, \ldots, 2^{C-1}, 2^C].$$

Uma vantagem do protocolo de Difusão Rápida (DR) é a sua possibilidade de integração com transmissões ao vivo, permitindo que novos usuários assistam a estas com um pequeno atraso em relação à transmissão original.

O protocolo DR também possui outra vantagem: permite servir vídeos a clientes sem *buffer* de armazenamento. Neste caso, o tempo máximo de espera sobe para metade da duração do vídeo escolhido. Na Figura 4.7, são assinalados por t_1 e t_2 os momentos de início de transmissões que não necessitam de armazenamento.

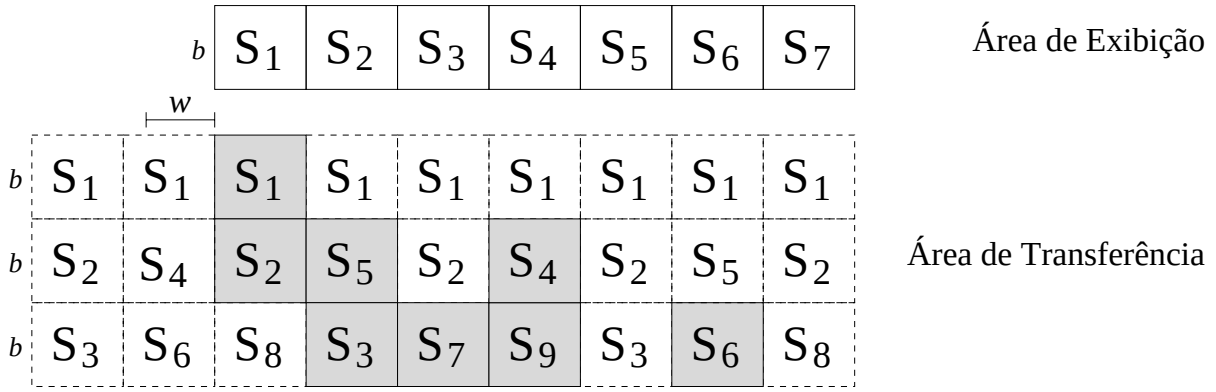

Figura 4.8: Mapa do Protocolo de Difusão em Pagode

4.9 Protocolo de Difusão em Pagode

Protocolos da família do Protocolo de Difusão em Pagode [34] (*Pagoda Broadcasting*) possuem características intermediárias aos protocolos da família de Difusão Piramidal e de Difusão Harmônica (seção 4.11): utilizam segmentos de mesmo tamanho, transmitidos em canais com a mesma largura de banda que a taxa de consumo do vídeo. A novidade introduzida pelo protocolo de Difusão em Pagode (PDPa) está na freqüência de transmissão dos segmentos, agora não mais transmitidos em canais exclusivos (Figura 4.8). Neste protocolo, cada segmento i é alocado sempre em um mesmo canal, de forma que seja transmitido no máximo a cada i fatias de tempo. O segmento S_1, portanto, ocupa exclusivamente o primeiro canal. Já o segundo canal transmite o segmento S_2 (a cada duas fatias de tempo) e os segmentos S_4 e S_5 (ambos transmitidos a cada quatro fatias de tempo). Por fim, o terceiro canal transmite o segmento S_3 com uma freqüência igual a $1/(3f)$, onde f corresponde a uma fatia de tempo. Cada um dos segmentos restantes S_6,S_7,S_8 e S_9 é transmitido a uma freqüência de $1/(6f)$. Em comparação com o Protocolo de Difusão Rápida, o PDPa é mais eficiente, pois consegue com a mesma largura de banda (três canais) dividir o vídeo em nove partes iguais, contra sete segmentos do Protocolo de Difusão Rápida.

Generalizando a política de alocação dos segmentos, cada par de fluxos consecutivos $(2k,2k+1)$ abriga $4z$ segmentos, dispostos de acordo com a Tabela 4.1. z é definido como o índice de um segmento S_z com menor número dentre os segmentos alocados para o par de fluxos citado, e é obtido através da seguinte função:

$$z(2k+1) = 2(5^{k-1}) \tag{4.3}$$

Segmentos	Fluxo	Freqüência de transmissão
S_z a $S_{3z/2-1}$	$2k$	$1/(zf)$
$S_{3z/2}$ a S_{2z-1}	$2k+1$	$2/(3zf)$
S_{2z} a S_{3z-1}	$2k$	$1/(2zf)$
S_{3z} a S_{5z-1}	$2k+1$	$1/(3zf)$

Tabela 4.1: Mapeamento de segmentos em canais do Protocolo de Difusão em Pagode

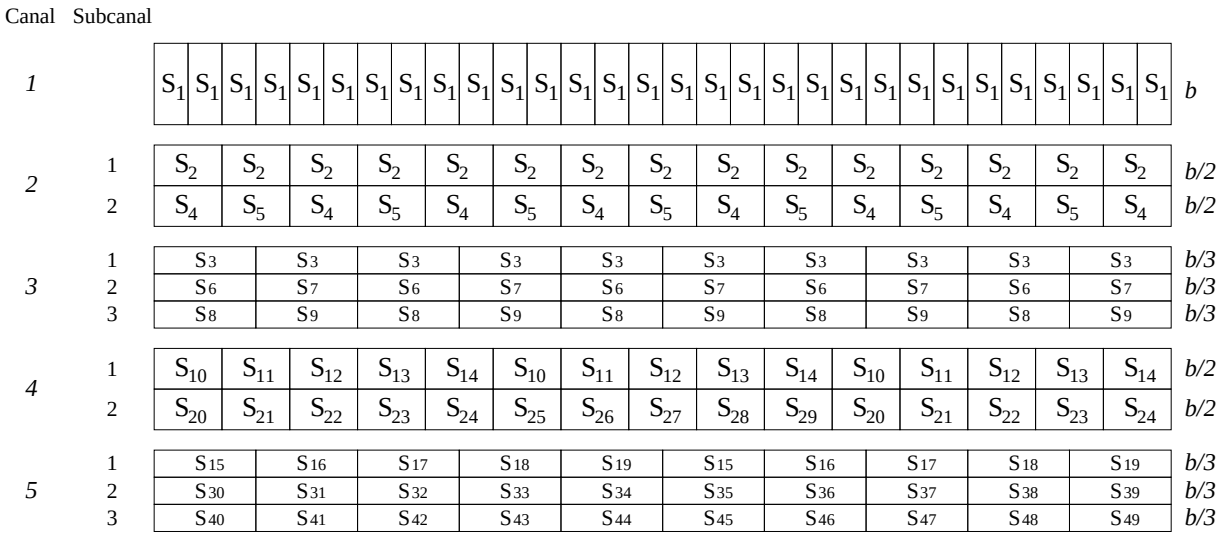

Figura 4.9: Divisão dos canais do Novo Protocolo de Difusão em Pagode

Ao tomar os canais quatro ($2k$) e cinco ($2k+1$) como exemplo, tem-se $z = 10$ e, portanto, os segmentos entre S_{11} e S_{14} transmitidos no quarto canal, a cada 10 fatias de tempo. Segmentos de S_{20} a S_{29} também são transmitidos no quarto canal, agora a cada 20 fatias de tempo. O quinto canal, por sua vez, abriga os segmentos entre S_{15} e S_{19} (transmitindo cada segmento em períodos de 15 em 15 fatias de tempo) e entre S_{30} e S_{49} (a cada 30 fatias de tempo).

O número total de segmentos $N(n)$ do Protocolo de Difusão em Pagode é dado por

$$N(n) = \begin{cases} 4(5^{k-1}) - 1 & n = 2k \\ 2(5^k) - 1 & n = 2k+1 \end{cases} \tag{4.4}$$

Como requisito de armazenamento do cliente, o PDPa exige o armazenamento de aproximadamente 46% do vídeo.

S_{10}	S_{11}	S_{12}	S_{13}	S_{14}
S_{17}	S_{16}	S_{15}	S_{18}	S_{19}

S_{10}	S_{12}	S_{14}	S_{16}	S_{19}
S_{11}	S_{13}	S_{15}	S_{17}	S_{20}
...	...	...	S_{18}	S_{21}

Figura 4.10: Matriz retangular utilizada para mapeamento de segmentos em canais para PDPa e NPDPa

4.10 Novo Protocolo de Difusão em Pagode

O Novo Protocolo de Difusão em Pagode[33] baseia-se em uma nova política de alocação dos segmentos entre os canais existentes. A idéia principal é baseada no fato de que um mapeamento de segmentos mais eficiente entre os canais disponíveis gera um aumento no número de segmentos, diminuindo assim o tempo de espera. Para permitir esta nova alocação, é utilizada uma matriz retangular de mapeamento. Nesta matriz, cada linha é composta por um determinado número de fatias de tempo transmitidas consecutivamente. Para efeito de comparação dos protcolos, a Figura 4.10 descreve um exemplo de otimização do mapeamento para o quarto canal do Protocolo de Difusão em Pagode (à esquerda), onde o alocamento é feito de forma a preencher horizontalmente toda a matriz. No caso do NPDPa (à direita), o preenchimento é feito por colunas, permitindo a adição de dois segmentos, S_{20} e S_{21}. A principal explicação para a melhor performance nesta alocação é que a periodicidade dos segmentos S_{16} a S_{19} diminui para 16 fatias de tempo entre cada transmissão, salvando então largura de banda suficiente para a adição dos dois segmentos. Este raciocínio é aplicado no NPDPa em todos canais.

Este mapeamento de segmentos em canais é na realidade executado através da multiplexação por tempo dos segmentos alocados em subcanais[32], ilustrados pela Figura 4.9. O primeiro canal transmite exclusivamente o primeiro segmento a uma largura de banda b, garantindo sua transmissão completa a cada fatia de tempo. O segundo canal é subdividido em dois subcanais de largura de banda igual a $b/2$ através de multiplexação de tempo. O primeiro subcanal, indicado na figura como subcanal 1, contém todas as fatias de tempo pares do segundo canal, transmitindo o segmento S_2 continuamente a cada duas fatias de tempo. O subcanal 2, por sua vez, difunde alternadamente os segmentos S_4 e S_5. O canal 3 é subdividido em três subcanais transmitidos a uma largura de banda equivalente a $b/3$. Assim, cada segmento é transmitido nestes subcanais em três

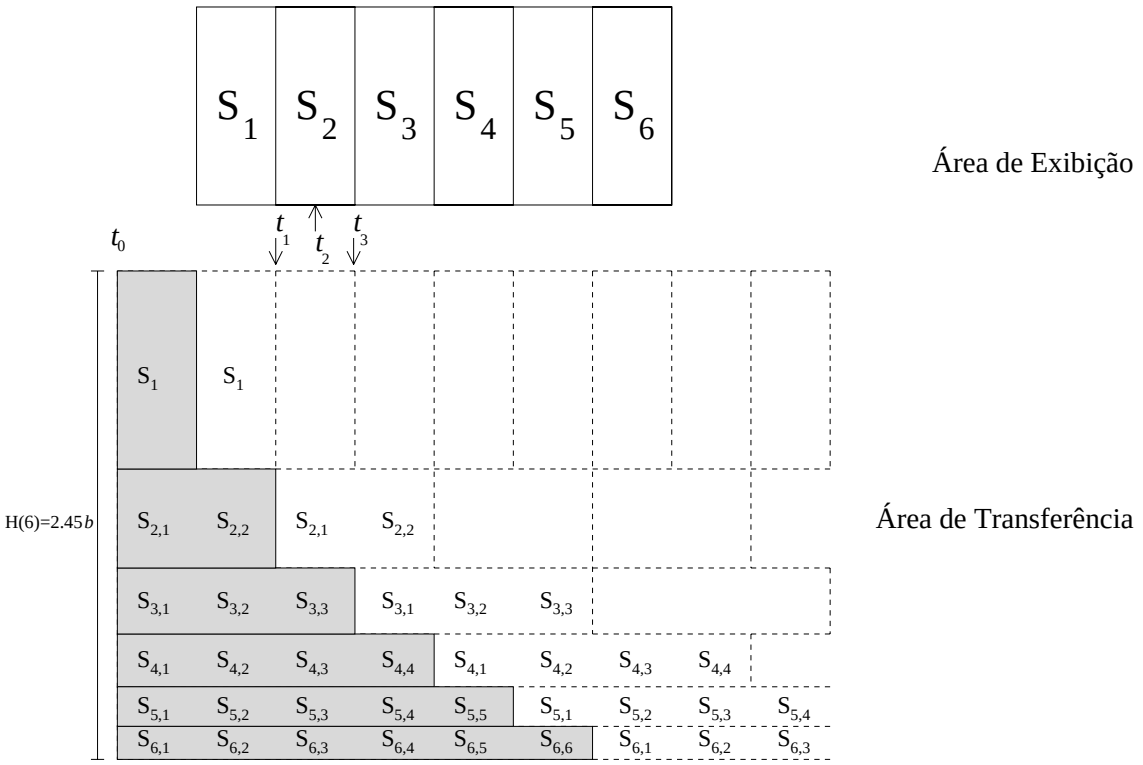

Figura 4.11: Mapa largura de banda-temporal do Protocolo de Difusão Harmônica

fatias de tempo. O subcanal 1 deve transmitir exclusivamente o segmento S_3, enquanto o subcanal 2 transmite repetidamente os segmentos S_6 e S_7. Da mesma forma, o subcanal 3 transmite S_8 e S_9. Novos canais sempre são adicionados aos pares: um canal com índice par (no exemplo, o canal 4, transmitindo os segmentos S_{10} a S_{14} no primeiro subcanal e S_{20} a S_{29} no segundo) é subdividido em dois subcanais e um canal com índice ímpar, em três subcanais (canal 5, S_{15} a S_{19} para o subcanal 1, S_{30} para o subcanal 2 e S_{40} a S_{49} para o 3). Desta forma, é possível mapear 49 segmentos em cinco canais lógicos no NPDPa.

4.11 Protocolo de Difusão Harmônica

O protocolo de Difusão Harmônica[25], ou *Harmonic Broadcasting*(PDH), segue um padrão diferente dos descritos anteriormente, e deu início a uma nova família de protocolos. O vídeo é dividido em n segmentos de tamanhos iguais, e a transmissão é realizada em canais lógicos de largura de banda decrescente, seguindo uma série harmônica. Cada segmento S_i é subdividido em mais i subsegmentos e transmitido em um canal exclusivo a uma

largura de banda igual a $\frac{b}{i}$. Desta forma, cada subsegmento leva uma fatia de tempo para ser transmitido, e todo o segmento é transmitido a cada i fatias de tempo. O cliente espera a ocorrência do início da transmissão do primeiro segmento e desde então recebe todos os segmentos ao mesmo tempo, requerendo uma largura de banda L proporcional à série harmônica:

$$L_{PDH(n)} = \sum_{i=1}^{n} \frac{b}{i} = bH(n)$$

onde $H(n)$ é o número harmônico de n.

Como a série harmônica em função do número de segmentos cresce lentamente[1], torna-se passível de um aumento substancial no número de segmentos (reduzindo-se assim o tempo de espera), sem com isso aumentar a largura de banda total necessária aos protocolos do tipo harmônico, o que torna vantajosa a utilização desta série. Deve-se salientar, entretanto, o compromisso entre o número de segmentos e a complexidade para gerenciá-los: o excesso de segmentos aumenta a complexidade de gerenciamento dos mesmos, tanto do lado do servidor quanto do cliente.

O PDH, entretanto, possui a grande desvantagem de não conseguir entregar segmentos no momento próprio em todos os casos. Na Figura 4.11, vê-se um exemplo ideal, onde a requisição é feita no momento t_0 e instantaneamente dá-se a recepção. Entretanto, se por ventura a recepção do cliente iniciar no momento t_1, pode-se perceber que a segunda parte do segmento S_2 ainda não foi recuperada para exibição no momento t_2 e só estaria disponível no momento t_3. O requisito de tempo de espera mínimo exato é equivalente a $\frac{(n-1)f}{n}$ unidades de tempo [36], aproximando-se a uma fatia de tempo para valores de n consideravelmente grandes.

4.12 Protocolo de Difusão Harmônica Cautelosa

Este protocolo é uma variação do Protocolo de Difusão Harmônica para evitar o problema de entrega dos segmentos em tempo hábil. Para tal, é adotada uma política de alocação de canais mais conservadora que seu protocolo antecessor. Como política conservadora, entenda-se por um aumento na largura de banda do canal (em relação ao canal correspondente do PDH) de forma que os dados sejam entregues mesmo no pior caso apontado no protocolo PDH.

O primeiro segmento (S_1) continua sendo transmitido em um canal exclusivo e utilizando a mesma largura de banda b que a taxa de consumo do vídeo. O segundo canal

[1] A série harmônica possui comportamento assintótico $O(\log n)$ [6]

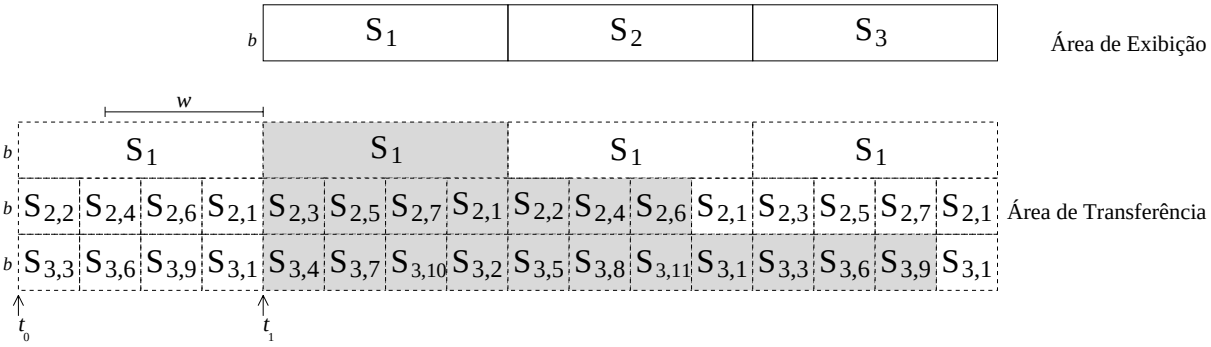

Figura 4.12: Mapa do Protocolo de Difusão Quase Harmônica (***m=4***)

muliplexa os segmentos S_2 e S_3, também a uma largura de banda b. Os $n-3$ canais restantes são transmitidos em canais lógicos exclusivos, com largura de banda decrescente. A diferença em relação à Difusão Harmônica é que cada segmento S_i é transmitido a uma largura de banda igual $\frac{b}{i-1}$. O Protocolo de Difusão Harmônica Cautelosa [36] consegue entregar a tempo todos os seus segmentos. Como os três primeiros segmentos são transmitidos com largura de banda b, não há como comprometer suas entregas. Cada i-ésimo segmento dentre os restantes é transmitido a cada $i-1$ fatias de tempo e, portanto, fazendo a entrega de todo o vídeo.

O Protocolo de Difusão Cautelosa requer uma largura de banda igual a:

$$\begin{aligned} L_{PDHC} &= 2b + \sum_{i=3}^{n-1} \frac{b}{i} \\ &= \frac{b}{2} + bH(n-1), \end{aligned}$$

ou seja, meio canal $(0,5b)$ a mais que o Protocolo de Difusão Harmônica.

4.13 Protocolo de Difusão Quase Harmônica

O Protocolo de Difusão Quase Harmônica (*Quasi-Harmonic Broadcasting*) [36], ou PDQH, é outro protocolo baseado no Protocolo de Difusão Harmônica e cujo objetivo principal está na correção da garantia de entrega dos segmentos a tempo. Ao contrário da Difusão Harmônica Cautelosa, permite que o cliente consuma os dados de um segmento enquanto os recebe.

A segmentação é idêntica à Difusão Harmônica, com n segmentos de tamanhos iguais. Da mesma forma, as fatias de tempo correspondem ao tempo para transmissão do seg-

mento S_1 a uma largura de banda b. O PDQH possui um parâmetro m, correspondente ao número de fragmentos transmitidos a cada fatia de tempo. Cada fatia de tempo é subdividida entre m subfatias de tempo. A Figura 4.12 ilustra um mapa do protocolo, com $m = 4$. Ainda na figura, pode-se ver uma fatia de tempo, marcada entre os instantes t_0 e t_1.

A diferença principal do protocolo reside na fragmentação (subsegmentação) dos segmentos. Excetuando-se o primeiro segmento (que não é fragmentado), cada segmento S_i é fragmentado em $im - 1$ partes. A alocação dos fragmentos para um segmento S_i é feita de forma que os primeiros $(i - 1)$ fragmentos sejam transmitidos sempre durante a última subfatia de cada fatia de tempo. Restam então $i(m - 1)$ fragmentos, que são transmitidos utilizando a seguinte regra: a k-ésima subfatia da fatia de tempo j é utilizada para transmitir o fragmento $(ik + j - 1) \bmod i(m - 1) + i$.

Resumindo, cada canal lógico no Protocolo de Difusão Quase Harmônica (PDQH) equivale a:

$$b_i = \begin{cases} 1 & i = 1 \\ \frac{m}{im-1} & i \geq 2 \end{cases} \tag{4.5}$$

No total, o requisito de largura de banda do protocolo de Difusão Quase Harmônica é igual a

$$L_{DQH} = b + \sum_{i=2}^{n} \frac{bm}{im - 1}$$

4.14 Protocolo de Difusão Poliharmônica

O protocolo de Difusão Poliharmônica (*Polyharmonic Broadcasting*)[35] foi concebido na tentativa de resolver o problema do tempo indevido de entrega de segmentos e melhorar a utilização de largura de banda. Da mesma família que o Protocolo PDH, o Protocolo de Difusão Poliharmônica (PDPH) também divide o vídeo em n segmentos de tamanho igual, transmitidos em canais exclusivos com largura de banda decrescente, tendo como base a série harmônica. O PDPH, entretanto, introduz duas grandes novidades em relação aos protocolos anteriores. A primeira delas reside no fato do cliente iniciar a recepção dos dados logo após o momento da requisição, e não após se dar o início da transmissão do primeiro segmento. Esta mudança implica na necessidade de se recuperar um segmento em sua totalidade antes de sua exibição, pois no pior caso o cliente inicia a recepção de um segmento instantes após o início de sua transmissão e deverá então recuperar os quadros

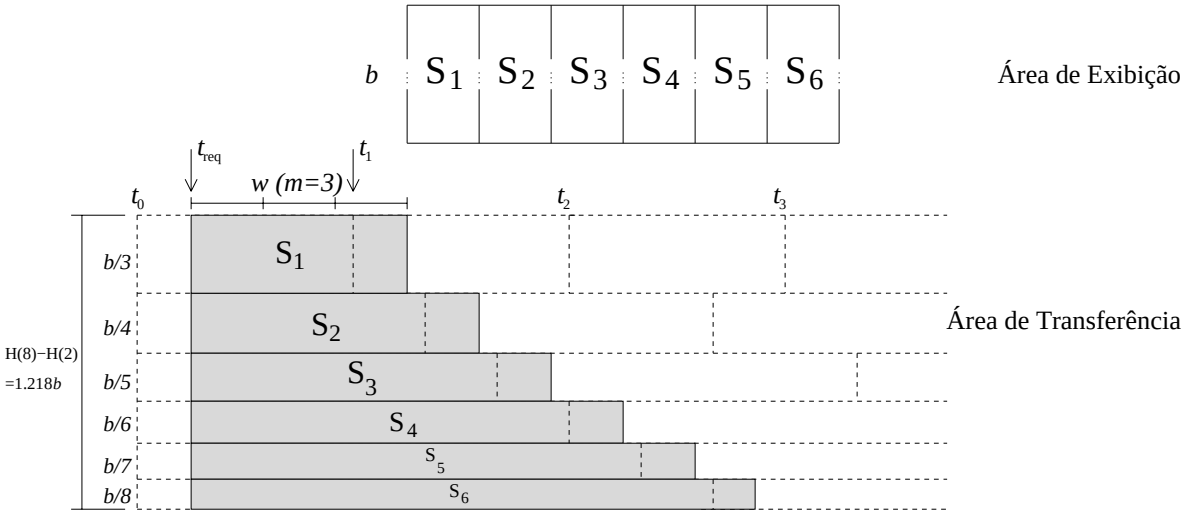

Figura 4.13: Mapa do Protocolo de Difusão Poliharmônica (***m=3***)

que faltaram após o início da próxima transmissão deste segmento (Figura 4.13). Este fato impõe também a segunda inovação: o protocolo de Difusão Poliharmônica faz com que os usuários sempre esperem um tempo fixo, correspondente à duração da transmissão do primeiro segmento. Um tempo de espera fixo é uma vantagem para o serviço de VoD, pois os usuários são sensíveis a diferentes tempos de espera.

Este protocolo possui um parâmetro adicional m, relacionado com a largura de banda do primeiro segmento, cujo valor é igual à taxa de consumo do vídeo b multiplicada pelo m-ésimo termo da série harmônica (ou seja, com o valor $\frac{b}{m}$). Os segmentos seguintes seguem normalmente a série harmônica: $\frac{b}{m+1}, \frac{b}{m+2}, \ldots, \frac{b}{m+n-1}$.

Além disso, cada segmento S_i tem o mesmo tamanho ($S_i = \frac{S}{n}, i = 1, 2, \ldots, n$) e é transmitido a uma largura de banda igual a $\frac{b}{m+i-1}$, isto é, são necessárias $m + i - 1$ fatias de tempo para recepção completa do vídeo.

Da mesma forma que no PDH, em PDPH considera-se uma fatia de tempo equivalente à n-ésima fração da duração do vídeo, S, ou seja,

$$d = \frac{S}{n} \tag{4.6}$$

A latência de exibição do PDPH equivale ao tempo de transferência do primeiro segmento, transmitido a $\frac{b}{m}$. Isto significa que levará m fatias de tempo para a recepção completa deste segmento:

$$w = md = m\frac{S}{n} \tag{4.7}$$

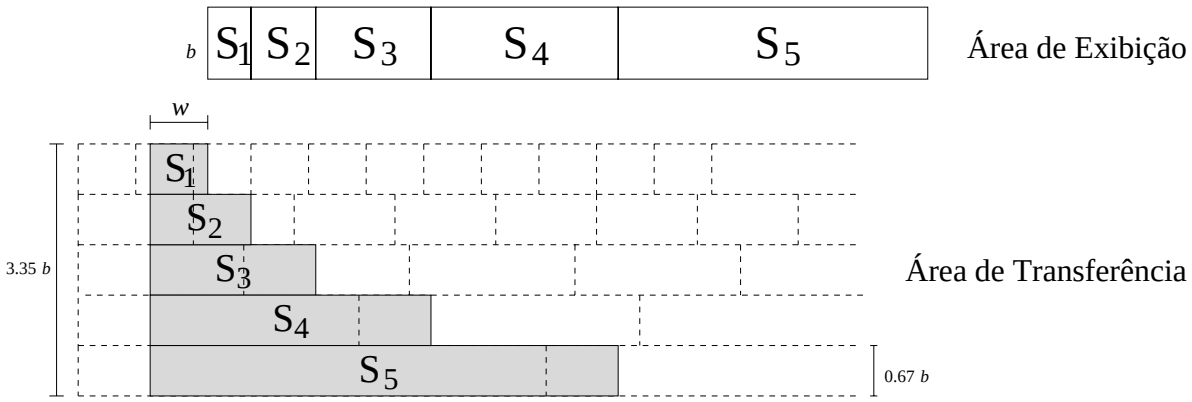

Figura 4.14: Mapa do GEBB ($n = 5$)

A largura de banda demandada pelo Protocolo PDPH é igual a

$$B_{PDPH}(m,n) = \sum_{i=1}^{n} b_i = b \sum_{i=1}^{n} \frac{1}{m+i-1} = b\Big[H(m+n-1) - H(m-1)\Big] \qquad (4.8)$$

onde $H(k)$ representa o número harmônico de k.

4.15 GEBB

O protocolo GEBB (*Greedy Equal Bandwidth Broadcating Protocol*, ou Protocolo de Difusão Guloso com [Canais de] Larguras de Banda Iguais) [22, 21] parte do princípio que a estrutura ótima de um protocolo de difusão decorre da mesma característica gulosa de recepção que o protocolo de Difusão Poliharmônica possui. Neste protocolo, o usuário deve iniciar a recepção de todos os canais (simultaneamente), a partir do momento de sua requisição. Semelhantemente ao PDPH, o GEBB também deve receber completamente todo segmento antes de poder exibí-lo.

Na Figura 4.14 mostra-se um exemplo da eficiência do GEBB em relação ao primeiro protocolo proposto, o Protocolo de Difusão Balanceada (PDB). Para obter os mesmos dez minutos de espera para um vídeo de duas horas, são necessários apenas cinco segmentos transmitidos em cinco canais de largura de banda igual $0,67b$ (perfazendo um total de $3.35b$), em contraste com os doze canais do PDB.

A obtenção da segmentação é feita através de um problema de otimização visando a minimização da utilização de largura de banda, dado por:

$$\min \sum_{i=1}^{n} b_i \tag{4.9}$$

Sujeito às seguintes restrições:

$$b_i\Big(w + \sum_{j=1}^{i-1} S_j\Big) = bS_i, \qquad i = 1, 2, \ldots, n \tag{4.10}$$

$$\sum_{i=1}^{n} S_i = S \tag{4.11}$$

$$S_i > 0, \;\; b_i > 0, \qquad i = 1, 2, \ldots, n \tag{4.12}$$

A primeira restrição (Equação 4.10) diz respeito à simultaneidade entre os momentos de término da recepção de um segmento e o de sua imediata exibição.

A segunda restrição (Equação 4.11), por sua vez, obriga o tamanho dos segmentos que compõem o vídeo devam ser dispostos de maneira que a soma da duração dos segmentos seja equivalente à duração de todo o vídeo.

A última restrição (4.12) define os limites inferiores e superiores das variáveis S_i e b_i.

A solução para este problema de otimização [22] pode ser obtida através da seguinte forma: inicialmente, atribui-se $b = 1$, fazendo com que os canais sejam obtidos em função da taxa de consumo do vídeo. A seguir, desenvolvendo-se a primeira restrição (Equação 4.10), tem-se:

$$b_i = \frac{S_i}{w + \sum_{j=1}^{i-1} S_j} \qquad i = 1, \ldots, n$$

$$b_i + 1 = \frac{S_i}{w + \sum_{j=1}^{i-1} S_j} + \frac{w + \sum_{j=1}^{i-1} S_j}{w + \sum_{j=1}^{i-1} S_j} \qquad i = 1, \ldots, n$$

$$b_i + 1 = \frac{w + \sum_{j=1}^{i} S_j}{w + \sum_{j=1}^{i-1} S_j} \qquad i = 1, \ldots, n$$

Multiplicando-se os n termos das quantidades $(1 + b_i)$, obtém-se

$$\prod_{i=1}^{n} (1 + b_i) = \frac{w + \sum_{j=1}^{n} S_j}{w} = \frac{S}{w} + 1 \tag{4.13}$$

O produto das n variáveis $(1 + b_i)$ é igual a uma constante $\frac{S}{w} + 1$. Para efeito de otimização, a função para minimização $\sum_{i=1}^{n} (b_i)$ é equivalente à função $\sum_{i=1}^{n} (b_i + 1)$. Ou seja:

$$min \sum_{i=1}^{K} b_i \equiv min \sum_{i=1}^{K} (1 + b_i) \tag{4.14}$$

Combinando as duas equações anteriores, tem-se que o resultado ótimo é obtido quando todas as variáveis b_i possuem o mesmo valor. Seja este valor ótimo chamado de $(1 + b^*)$. Tem-se:

$$b_i = b^* = \sqrt[n]{\frac{S}{w} + 1} - 1 \tag{4.15}$$

e

$$S_i = wb^*(1 + b^*)^{i-1} \tag{4.16}$$

como solução final do problema de otimização. A partir da equação 4.15, obtém-se a largura de banda total exigida pelo protocolo:

$$B_{GEBB} = \sum_{i=1}^{n} b_i = n\left(\sqrt[n]{\frac{S}{w} + 1} - 1 \right) \tag{4.17}$$

Embora não haja nenhuma restrição quanto ao tipo de segmentação que o protocolo deva possuir, a solução ótima é obtida quando são utilizados tamanhos de segmentos crescentes e mesma largura de banda para cada canal. A progressão utilizada na segmentação do protocolo GEBB corresponde a uma inovação em relação aos protocolos anteriores, uma vez que esse utiliza a latência de exibição e o tamanho total do vídeo como parâmetros da progressão do tamaho dos segmentos.

Este é o primeiro protocolo que parte de um valor de espera arbitrário, para a obtenção de valores de largura de banda por canal e de tamanho dos segmentos necessários para a geração de um calendário de recepção. Além disso, pode-se utilizar outra abordagem para saber a fração de tempo de espera a partir da largura de banda disponível, como será apresentado na Seção 5.6 para o caso do protocolo PDPH-LBU com um conjunto de canais.

4.16 Pré-carregamento parcial de segmentos (*partial preload*)

Uma técnica que permite a eliminação da latência de exibição em protocolos de difusão periódica é o chamado pré-carregamento parcial de segmentos (*partial preload*) [37]. Esta técnica consiste no aproveitamento do *buffer* da *set-top-box* para armazenar os prefixos dos vídeos providos por difusão periódica (notadamente, os mais freqüentemente acessados, uma vez que a chance de requisição de um destes filmes é maior que os demais). Outra vantagem desta técnica é que possibilita a utilização de vídeos comprimidos, pois a porção previamente carregada corresponde a um *buffer*, que compensa variação de quadros recebidos em uma sessão.

Independentemente da alternativa utilizada para sanar a falta do prefixo, a implementação do pré-carregamento parcial exige a verificação de duas situações especiais, as quais o prefixo pode estar indisponível no momento, a saber:

Escolha de um vídeo não popular Neste caso, o prefixo não estará armazenado, nem tampouco o vídeo será transmitido via difusão, sendo servido individualmente ou utilizando-se uma técnica de difusão seletiva.

Mudança no conjunto de vídeos Qualquer mudança no conjunto de vídeos transmitidos — seja pela inclusão de novos vídeos ou mudança na ordem de requisição dos vídeos — precisa ser atualizada pelas *set-top-boxes*. É necessário então um mecanismo de aviso para sincronização destes conjuntos, de forma que o tempo de transição dos conjuntos de vídeo (no qual novos vídeos ainda estejam indisponíveis, enquanto vídeos antigos perdem sua disponibilidade) seja minimizado.

Requisições seqüenciais São duas as principais situações possíveis para a indisponibilidade do prefixo. A primeira delas ocorre quando o cliente desejar assistir dois ou mais vídeos consecutivamente. Neste caso, os prefixos dos vídeos exibidos posteriormente serão sobrescritos conforme o andamento da recepção do primeiro vídeo. A segunda situação possível ocorre quando o cliente interrompe a exibição de um vídeo para imediatamente solicitar outro. Nesta situação, o pior caso ocorre quando todo o *buffer* contendo os prefixos foi sobrescrito para armazenamento dos segmentos do vídeo interrompido.

Uma primeira alternativa para solucionar este problema exige o dobro da capacidade necessária para a *set-top-box*: metade da capacidade armazena os prefixos,

enquanto a outra metade é utilizada para o vídeo requisitado no momento. Desta forma, torna-se desnecessário sobrescrever os prefixos enquanto chegam novos dados. Uma vantagem para o servidor é que esta abordagem não requer largura de banda adicional. Neste caso, entretanto, a *set-top-box* pode ter seu custo de armazenamento aumentado drasticamente. Outras duas alternativas utilizando largura de banda foram propostas. A primeira delas baseia-se na transmissão do prefixo sob demanda. A outra alternativa é a transmissão periódica destes prefixos. Utilizando esta opção, pode-se também aumentar a freqüência de transmissão dos prefixos: ao invés do servidor alocar um canal para transmitir d minutos dos V vídeos, utiliza-se V canais transmitindo o prefixo de cada vídeo em um canal individual.

4.17 Protocolos de Difusão Periódica Otimamente Estruturados

Em [21] foram destacadas três regras principais para os protocolos de difusão periódica, nas quais seu cumprimento acarreta em uma utilização mais racional da largura de banda. São elas:

Regra 1 Não possuir redundância na transmissão de um segmento — contado a partir do momento da requisição do usuário até o momento de início da exibição do mesmo segmento — em uma mesma transmissão. Protocolos que violam esta regra assumem a necessidade de iniciação da recepção dos segmentos somente após o início da transmissão do primeiro segmento, mas neste caso a porção de dados que deixa de ser aproveitada decorrente desta abordagem acaba por torná-los ineficientes. Dentre estes protocolos, incluem-se o Protocolo de Difusão Piramidal, o de Difusão Piramidal Baseado em Permutações, o Difusão de Arranha-céu e o de Difusão Harmônica Cautelosa.

Regra 2 Protocolos eficientes de difusão periódica não podem manter porções de dados desnecessárias em nenhum de seus ciclos. Se um protocolo o fizer, significa que a área útil de difusão é intercalada em algum momento com porções de dados sem utilidade para aquela recepção específica (vide Figuras 4.8 e 4.12). Os protocolos de Difusão Quase Harmônica, Difusão em Pagode e o Novo Protocolo de Difusão em Pagode violam esta regra.

Regra 3 Todos os segmentos devem ser entregues a tempo, sejam eles recebidos durante a própria exibição ou armazenados anteriormente na *set-top-box*. Esta é uma característica necessária para o funcionamento de qualquer protocolo de difusão periódica. Apenas o Protocolo de Difusão Harmônica, sem a correção devida do tempo de espera (Seção 4.11) viola esta regra.

Protocolos de difusão periódica que seguem estas regras são candidatos a minimizar o desperdício de banda passante e por este motivo são denominados otimamente estruturados (*optimally-structured*).

Dentre todos os protocolos de difusão periódica citados anteriormente, os únicos protocolos considerados otimamente estruturados são o PDPH e o GEBB.

4.18 Síntese do Capítulo

Protocolos de difusão periódica são ideais para vídeos populares, uma vez que requerem largura de banda constante independentemente do número de usuários, embora exijam latência de exibição, variável de acordo com o protocolo utilizado e os recursos disponíveis.

Os protocolos de difusão periódica são divididos em três famílias: a primeira é a família piramidal, onde os segmentos são transmitidos a uma largura de banda constante com segmentos de tamanho crescente. Outra é a família harmônica, onde os segmentos possuem tamanho igual e largura de banda decrescente. Por fim, a terceira família é um híbrido das anteriores, sendo a segmentação igual, assim como a largura de banda.

Protocolos otimamente estruturados são protocolos de difusão periódica cuja política de recepção gulosa não esperam a ocorrência do início do segmento para receberem dados. Desta forma, não desperdiçam a banda passante e são candidatos a se aproximar do limite teórico. Destacam-se nesta categoria os protocolos PDPH e GEBB.

Capítulo 5

Protocolos de Difusão Periódica Sujeitos a Limitação de Banda passante

Dentre os principais recursos a serem gerenciados em sistemas de VoD, pode-se destacar a largura de banda do servidor, a do cliente, a capacidade de armazenamento do servidor e do cliente, a capacidade de E/S dos dispositivos de armazenamento, bem como a capacidade de processamento da *set-top-box*.

A largura de banda é um dos principais recursos que devem ser levados em conta para a implantação de um sistema de VoD. Quanto menor for a demanda de banda passante, mais usuários poderão ser agregados ao sistema. Entretanto, a maioria dos protocolos não foi concebida para suportar clientes com limitações de largura de banda.

Todos os protocolos apresentados no Capítulo 4, excetuando-se o Protocolo de Difusão Balanceada e o Protocolo de Difusão Arranha-céu, requerem uma *set-top-box* capaz de receber entre cinco a dez canais simultâneos. Os protocolos otimamente estruturados, por exemplo, requerem que todos os canais sejam recebidos ao mesmo tempo, não possuindo nenhum tipo de adaptação ou adequação no caso da largura de banda do cliente ser insuficiente para tal recepção.

A motivação principal para o presente trabalho é estender os protocolos otimamente estruturados, de forma a permitir o suporte a um número maior de usuários no sistema, quando os usuários possuem limitação de banda passante.

Este capítulo apresenta duas adaptações, já existentes na literatura, que comportam clientes com banda passante limitada: o Protocolo de Difusão Rápida (Seção 5.2) e o

Novo Protocolo de Difusão em Pagode (Seção 5.3).

As duas seções seguintes introduzem a principal contribuição desta dissertação: duas extensões de protocolos otimamente estruturados, capazes de suportar clientes com banda passante limitada: o Protocolo de Difusão Poliharmônica com Limitação de Banda do Usuário (PDPH-LBU, na Seção 5.5) e o *Greedy Equal Bandwidth Brodcasting* com Limitação de Banda do Usuário (GEBB-LBU, na Seção 5.6).

5.1 Protocolos da famíla Pagode com limitação de banda do usuário

Em [38], discute-se um novo método para adaptação dos protocolos com o objetivo de transpor a limitação de largura de banda do cliente, adaptando os protocolos de Difusão Rápida e o Novo Protocolo de Difusão em Pagode. Ao contrário dos protocolos que recuperam dados de todos os C canais, este método baseia-se na recepção dos dados utilizando não mais do que k canais simultâneos (considerando canais de largura de banda b, isto é, clientes com uma largura de banda limitada a kb). Os $C - k$ canais restantes passam a ter sua recepção adiada pelos clientes e iniciada imediatemente depois da liberação de canais que já tenham sido utilizados. Assim, o cliente recebe segmentos do canal $k + 1$ somente após a recepção completa dos segmentos do primeiro canal; do canal $k + 2$ após a recepção do canal de número dois, e assim sucessivamente, até o C-ésimo canal, que terá sua recepção iniciada após a recepção total do $(C - k)$-ésimo canal. Vale lembrar que cada canal transmitido possui a mesma banda passante que a taxa de consumo do vídeo (b).

O critério de corretude de protocolos de difusão periódica da família Pagode baseia-se na repetição de um segmento S_i a cada i fatias de tempo. Esta restrição também se aplica para os protocolos de difusão periódica com restrição de largura de banda do cliente, para os primeiros k canais. A principal diferença reside nos canais restantes, onde a recepção é atrasada e, portanto, permitindo um menor número de segmentos para cada canal.

Em qualquer protocolo com limitação de largura de banda para recepção, a diminuição desta largura de banda acarreta em um aumento da latência de exibição, como nos casos dos protocolos PDR e NPDPa. Para estes últimos, isto fica mais evidente devido à diminuição do número total de segmentos, cuja conseqüência imediata é justamente o aumento da latência de exibição.

5.2 Protocolo de Difusão Rápida Limitada

O Protocolo de Difusão Rápida permite duas abordagens de divisão de segmentos (como visto na Seção 4.8). Para a execução do método supracitado, utiliza-se no PDR a abordagem de divisão do vídeo em segmentos iguais, como nos protocolos da família Pagoda.

Como a largura de banda do cliente está limitada a três canais, a *set-top-box* não poderá armazenar dados do quarto canal até que um dos três primeiros canais seja liberado, ou seja, efetivamente o primeiro canal que é liberado após a primeira fatia de tempo (logo após a recepção do primeiro segmento). Este atraso em uma fatia de tempo na recepção do quarto canal obriga a mudança da freqüência de repetição de um segmento de oito para sete fatias de tempo.

	Difusão Rápida		DR Limitada ($k = 3$)			DR Limitada ($k = 4$)		
Canal	Segmentos	NS	Atraso	Segmentos	NS	Atraso	Segmentos	NS
1	S_1	1	0	S_1	1	0	S_1	1
2	S_2 a S_3	2	0	S_2 a S_3	2	0	S_2 a S_3	2
3	S_4 a S_7	4	0	S_4 a S_7	4	0	S_4 a S_7	4
4	S_8 a S_{15}	8	1	S_8 a S_{14}	7	0	S_8 a S_{15}	7
5	S_{16} a S_{31}	16	2	S_{15} a S_{27}	13	1	S_{16} a S_{30}	15
6	S_{32} a S_{63}	32	4	S_{28} a S_{51}	24	2	S_{31} a S_{59}	29
7	S_{64} a S_{127}	64	8	S_{52} a S_{95}	44	4	S_{60} a S_{115}	56
8	S_{128} a S_{255}	128	15	S_{96} a S_{176}	81	8	S_{116} a S_{223}	108
9	S_{256} a S_{511}	256	28	S_{177} a S_{325}	149	16	S_{224} a S_{431}	208
10	S_{512} a S_{1023}	512	52	S_{326} a S_{599}	274	31	S_{432} a S_{832}	401

Tabela 5.1: Quadro comparativo da disposição dos segmentos utilizados nos protocolos de Difusão Rápida Limitada ($k = 3$ e $k = 4$)

A Tabela 5.1 exemplifica a transmissão do Protocolo de Difusão Rápida Limitada a três e quatro canais simultâneos, indicando o atraso (em fatias de tempo), os segmentos alocados e o número de segmentos (indicados pela coluna NS) alocados em cada canal do servidor. Para dez canais, por exemplo, um cliente com limitação do cliente em três canais pode dividir o vídeo em 599 segmentos, o que corresponde a um tempo de espera equivalente a 0,15% da duração do vídeo. Pode-se perceber que os atrasos vão se acumulando nos canais subseqüentes, o que aumenta a diferença no número total de segmentos. Por exemplo, o aumento na demanda de seis para dez canais no servidor causa a redução do número de segmentos de 19,05% para 41,45%.

S_{10}	S_{13}	S_{17}
S_{11}	S_{14}	S_{18}
S_{12}	S_{15}	S_{19}
...	S_{16}	S_{20}
...	...	S_{21}

Figura 5.1: Matriz retangular utilizada para mapeamento de segmentos em canais para o NPDPa limitado a três canais

Para clientes com quatro canais, utiliza-se o mesmo procedimento na alocação dos segmentos, respeitando-se agora a nova limitação de recepção de até quatro canais simultâneos. A Tabela 5.1 mostra a diferença entre a versão sem restrições e a versão limitada do protocolo de difusão rápida. Pode-se constatar também na tabela o ganho crescente em número de segmentos alocados por canal em relação à versão com a limitação de três canais, uma vez que a restrição imposta é menor.

Uma característica do Protocolo de Difusão Rápida é perdida em relação ao seu similar sem restrições: a opção de suporte a usuários sem capacidade de armazenamento (*buffer*). Devido ao custo decrescente de dispositivos de armazenamento, entretanto, esta característica perde a relevância ao longo do tempo.

5.3 Novo Protocolo de Difusão em Pagode Limitada

Para o Protocolo de Difusão Novo Pagode, a adaptação se dá de forma equivalente ao de Difusão Rápida: canais posteriores são liberados para recepção à medida em que os primeiros são recuperados. Ajustes de periodicidade também devem ser realizados. Da mesma forma que seu equivalente sem restrições, utiliza-se a matriz retangular (Seção 4.10) para remapear os segmentos nos canais.

A transmissão dos três primeiros canais é feita de forma idêntica ao protocolo original. Já no quarto canal do servidor sofre a primeira fatia de tempo de atraso, já que o primeiro canal é liberado logo após a recepção do primeiro segmento. A Figura 5.1 ilustra o mapeamento do quarto canal, ligeiramente diferente da Seção 4.10. Neste caso, os segmentos são agrupados em três colunas, de forma que os segmentos S_{10} a S_{12} sejam repetidos a

cada 9 fatias de tempo, os segmentos S_{13} a S_{16} a cada 12 fatias de tempo, e, os segmentos entre S_{17} e S_{21} repetidos a cada 15 fatias de tempo.

	NPDPa Limitada ($k = 3$)			**NPDPa Limitada ($k = 4$)**		
Canal	**Atraso**	**Segmentos**	**NS**	**Atraso**	**Segmentos**	**NS**
1	0	S_1	1	0	S_1	1
2	0	S_2 , S_4 e S_5	(3)	0	S_2 , S_4 , S_8 e S_9	(4)
3	0	S_3 , S_6 a S_9	5	0	S_3 , S_6 e S_7, S_{12} a S_{14}, S_{25} e S_{26}	(8)
4	1	S_{10} a S_{21}	12	0	S_5, S_{10} e S_{11} S_{15} a S_{24}	(13)
5	4	S_{22} a S_{46}	25	1	S_{27} a S_{62}	36
6	6	S_{47} a S_{107}	61	8	S_{63} a S_{140}	78
7	16	S_{108} a S_{249}	142	24	S_{141} a S_{318}	178
8	40	S_{250} a S_{582}	333	24	S_{319} a S_{791}	473

Tabela 5.2: Quadro comparativo da disposição dos segmentos utilizados no Novo Protocolo de Difusão em Pagode, limitado a três e quatro canais

A Tabela 5.2 exibe uma comparação entre as duas versões do NPDPa com limitação de banda passante. A coluna NS (número de segmentos multiplexados em um canal) indica quantos segmentos são transmitidos no canal correspondente. Os itens entre parênteses, na coluna de número de segmentos indica que não há configuração definida para o número de canais indicado (exemplificando, não há configuração para o NPDPa limitado a $k = 4$ utilizando somente três canais de transmissão, uma vez que os segmentos S_5, S_{10} e S_{11} são transmitidos somente no quarto canal). O número total (acumulado) de segmentos até um determinado canal pode ser obtido na tabela através do maior índice do segmento contido neste canal. O Protocolo de Difusão Rápida utiliza 255 segmentos com oito canais, enquanto suas versões com limitação de banda do usuário utilizam 176 e 223 segmentos para $k = 3$ e $k = 4$, respectivamente.

Para a limitação de clientes com quatro canais, utiliza-se a mesma metodologia que o caso com três canais, fazendo-se o mapeamento dos segmentos através da matriz retangular. Os resultados estão sumarizados na Tabela 5.2, onde mostra-se o ganho evidente em relação à versão com limitação de três canais: utilizando-se oito canais do servidor, por exemplo, a versão limitada a quatro canais possui 42% mais segmentos que a versão

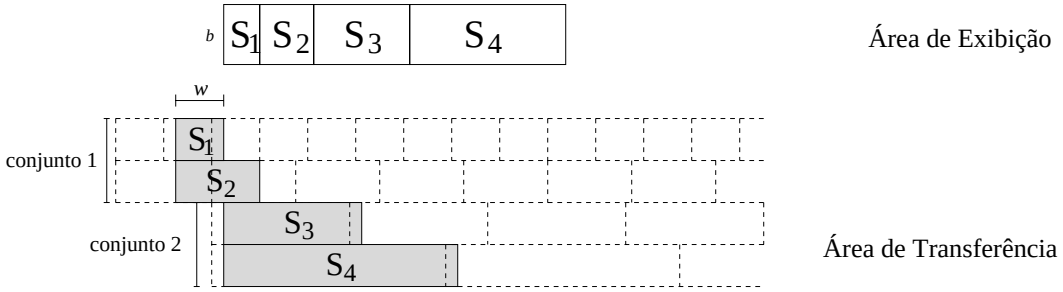

Figura 5.2: Conjuntos de canais

limitada a três canais simultâneos, o que melhora sensivelmente a latência de exibição.

5.4 Protocolos otimamente estruturados com limitação de banda do usuário

Um dos principais objetivos de qualquer protocolo de difusão periódica é a obtenção da menor latência de exibição possível. Como a latência está diretamente associada ao tamanho do primeiro segmento, torna-se possível reduzí-la através de uma maior segmentação do vídeo, reduzindo o prefixo e, por conseqüência a latência. Entretanto, o aumento no número de segmentos acaba por demandar uma largura de banda extra para acomodá-los. Nos casos dos protocolos da família Pagode, esta banda extra surge ao alocarmos mais um canal para acomodar os segmentos recém-criados. Nas famílias restantes (incluindo os protocolos otimamente estruturados), cria-se um canal lógico para cada segmento criado. Por outro lado, havendo limites na largura de banda disponível (tanto no lado do cliente quanto do servidor) e na segmentação (para não aumentar demasiadamente a complexidade da *set-top-box*), surgem limitações também na latência mínima obtida. Ao utilizar somente os três primeiros canais do PDR, por exemplo, permite-se apenas a utilização de sete segmentos, o equivalente a um tempo de espera igual a 14,28% do tempo total do vídeo. Para este protocolo, qualquer latência inferior a este valor é impossível de ser alcançada mediante esta restrição.

Um conjunto de canais é definido aqui como um grupo de canais adjacentes que possui início de recepção no mesmo instante. Na Figura 5.2, por exemplo, os canais dos segmentos S_1 e S_2 formam um conjunto, enquanto os canais que transmitem S_3 e S_4 formam outro conjunto com um segmento de atraso (uma vez que o início da recepção se

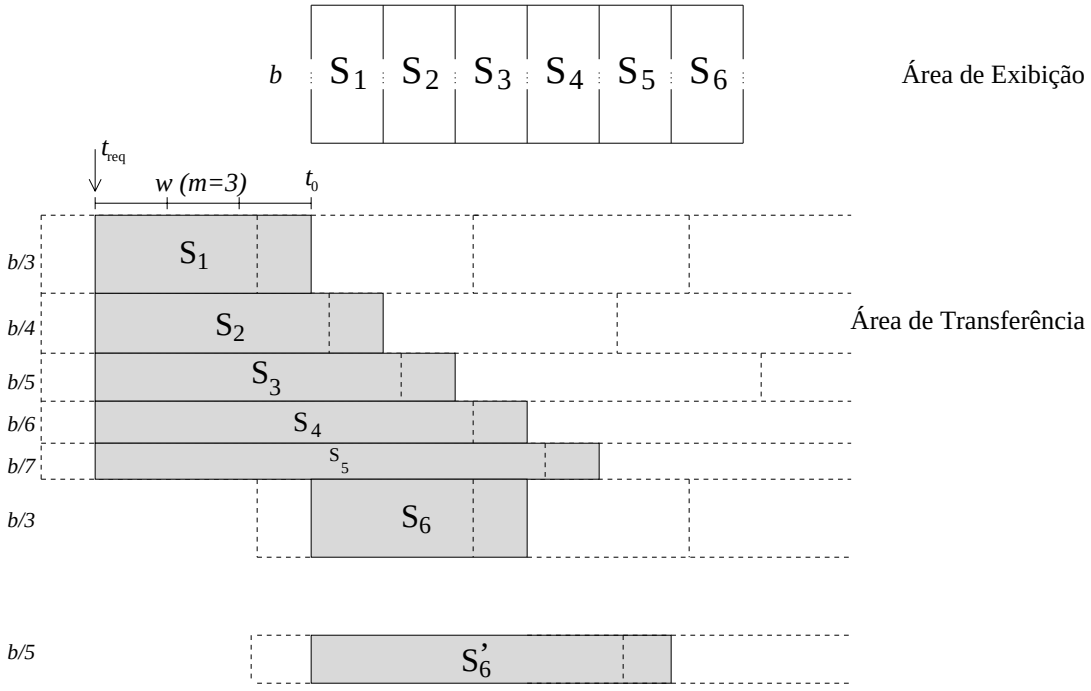

Figura 5.3: Perda de eficiência na troca de canais

dá somente após o término da recepção completa do primeiro segmento) em relação ao primeiro conjunto. Ao adicionar os canais em conjuntos, pode-se vê-los separadamente, permitindo a utilização de expressões mais simples, utilizadas nas soluções dos protocolos originais que possuem somente um conjunto de canais.

A abordagem usada para os protocolos PDR e NPDPa limitados, onde o término de utilização de um canal implica automaticamente na utilização de outro pode não ser adaptável para protocolos da família harmônica, como por exemplo o de Difusão Poli-harmônica. Isto se deve ao fato deste utilizar larguras de banda diferentes para cada canal. Considere o exemplo ilustrado na Figura 5.3: no instante t_0, há o término da recepção do primeiro canal. Neste momento, a substituição imediata por um canal equivalente não é o mais indicado. Pode-se optar pelo fluxo S'_6, com um ganho de $0,133b$ de banda para o servidor. Pretende-se então utilizar uma nova abordagem, incluindo conjuntos inteiros de canais com atraso e não mais canais individuais.

Outras configurações de calendários ainda poderiam ser propostas para atingir latências de exibição melhores com as mesmas restrições impostas. Comparando com o exemplo anterior, pode-se esperar a recepção do segmento S_6 até o término da recepção de S_2.

Neste momento, restam ao cliente $b/4$ de banda disponível, possibilitando a modificação da segmentação para comportar um segmento a mais e, conseqüentemente, diminui-se a latência de exibição. Note que esta opção exige um aumento na demanda de largura de banda do servidor, mas em compensação diminui-se a latência de exibição. Devido à vasta gama de possibilidades na configuração dos vários parâmetros modificadores do calendário (como por exemplo o tamanho de cada segmento, a largura de banda dos canais, o tempo de atraso dos conjuntos de canais e o número de segmentos de cada conjunto), utiliza-se um problema de otimização para determinar a configuração ótima destes parâmetros.

A escolha de protocolos otimamente estruturados foi feita devido ao fato destes possuírem uma utilização mais racional da banda disponível, o que leva a um melhor desempenho em relação ao restante dos protocolos. A idéia principal é que a restrição de largura de banda do usuário não impeça a utilização racional da banda passante do servidor. Nesta seção, são introduzidos dois novos protocolos, resultados da extensões de protocolos otimamente estruturados. O primeiro deles, desenvolvido a partir do PDPH (Seção 4.14), é o Protocolo de Difusão Poliharmônica com Limitação de Banda do Usuário (PDPH-LBU). O segundo protocolo é uma extensão do GEBB (Seção 4.15) e é denominado *Greedy Equal Bandwidth Broadcasting* com Limitação de Banda Passante (GEBB-LBU).

Para a exposição destes protocolos, descritos nas seções seguintes, optou-se pela seguinte organização: primeiro obtém-se os resultados para somente um conjunto de canais, abordando cada uma das funções objetivo.

Só a partir daí o problema é estendido, através da divisão dos canais em conjuntos. O servidor transmite todos os canais de um determinado conjunto com um atraso em relação aos canais do conjunto anterior, para que um cliente possa receber cada conjunto utilizando sua capacidade máxima disponível. Com esta visão, os problemas de otimização para apenas um conjunto de canais são generalizados e resolvidos através de avaliação de todas as possibilidades ou algoritmos genéticos.

5.5 O protocolo PDPH-LBU

Nesta seção, apresenta-se o protocolo PDPH sob limitação de banda passante no cliente. O objetivo é obter um protocolo mais próximo de um otimamente estruturado, minimizando a demanda de banda passante durante a transmissão dos fluxos de vídeo.

A minimização da demanda de largura de banda do servidor (denotada pela função B) é um objetivo já utilizado pelo protocolo GEBB e busca otimizar um dos recursos mais importantes: a largura de banda do servidor, uma vez que seu ganho possibilita um

aumento no número de vídeos populares servidos pelo sistema de VoD.

Para a obtenção deste objetivo, basta somar a largura de banda de cada canal lógico alocado para o protocolo, ou seja:

$$B = \sum_{i=1}^{n} b_i \tag{5.1}$$

No PDPH, obtém-se a largura de banda demandada por cada canal através dos parâmetros m e n. O primeiro canal requer uma banda igual a $\frac{b}{m}$, o segundo requer $\frac{b}{m+1}$ e a redução segue até o último canal, que demanda $\frac{b}{m+n-1}$. De (5.1), obtém-se:

$$B = \frac{b}{m} + \frac{b}{m+1} + \cdots + \frac{b}{m+n-1}$$

$$= b\left[H(m+n-1) - H(m-1)\right], \tag{5.2}$$

onde $H(n)$ representa o número harmônico de n.

Dado que b é um termo multiplicativo, pode-se desprezar este termo na função objetivo. Desta forma, o resultado obtido em um problema de otimização que se utiliza desta função é dado em múltiplos da taxa nominal do vídeo. Além disso, a função torna-se dependente apenas dos parâmetros m e n, e é dada por:

$$B = \left[H(m+n-1) - H(m-1)\right] \tag{5.3}$$

5.5.1 Formulação do problema para um único conjunto de canais

Utilizando-se o PDPH-LBU com somente um conjunto de canais, nem sempre se obtém soluções viáveis, uma vez que as mesmas são arbitrárias e podem restringir o problema de tal forma que não haja segmentação factível. A sua vantagem, entretanto, reside na simplicidade, sendo assim, ideal para exposição inicial do problema de otimização completo.

Como parâmetros do problema, foram considerados a latência de exibição w, a qual os clientes são submetidos, e um parâmetro, k, correspondente à razão entre a largura de banda disponível do cliente e a largura de banda utilizada para a exibição contínua do vídeo. Por exemplo, um valor $k = 3$ limita o usuário a uma recepção máxima de $3b$ e é equivalente à restrição de três canais imposta ao PDR e ao NPDPa (Seção 5.2).

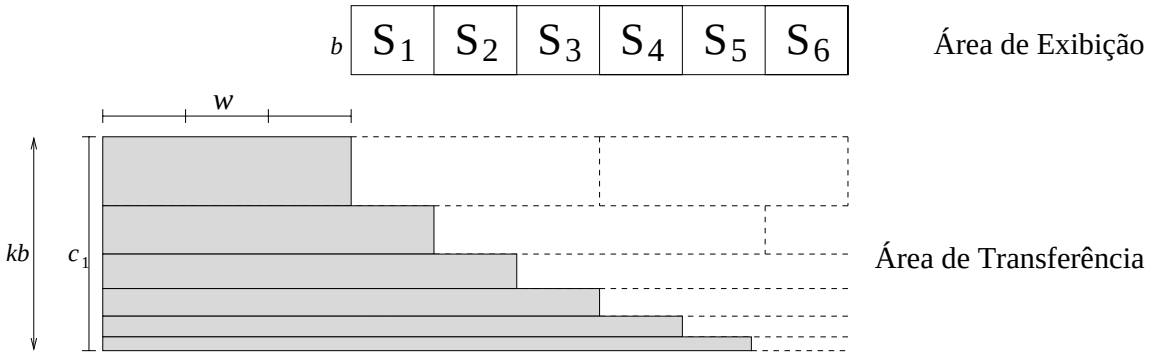

Figura 5.4: Mapa do PDPH-LBU para um conjunto de canais

A Figura 5.4 ilustra o mapa de difusão do PDPH com apenas um conjunto, c_1 composto por seis canais.

O problema de otimização associado ao PDPH com restrições, apresentado aqui com a denominação Protocolo de Difusão Poliharmônica com Limitação de Banda do Usuário, ou PDPH-LBU (*Polyharmonic Broadcasting with User Bandwidth Limit* [51]), é dado por:

$$\min \ [H(m+n-1) - H(m-1)] \tag{5.4}$$

Sujeito às seguintes restrições:

$$H(m+n-1) - H(m-1) \leq k \tag{5.5}$$

$$w \geq m\frac{S}{n} \tag{5.6}$$

$$1 \leq m \leq m_{max} \tag{5.7}$$

$$m\frac{S}{w} \leq n \leq n_{max} \tag{5.8}$$

onde m e n são inteiros, correspondem aos mesmos parâmetros utilizados no PDPH e representam, respectivamente, o termo inicial da série harmônica e o número de segmentos ao qual o vídeo é submetido. m_{max} e n_{max} definem o valor máximo destas duas variáveis . O parâmetro n_{max} define a granularidade da segmentação, acarretando em um compromisso entre desempenho e computabilidade. Se por um lado, uma segmentação menor traz um desempenho mais baixo (através de uma latência de exibição maior e, por vezes, também

uma maior demanda de banda passante), por outro lado uma grande segmentação acarreta em uma complexidade computacional maior para o gerenciamento destes segmentos, tanto na *set-top-box* quanto no servidor.

A primeira restrição (Inequação 5.5) trata do ponto central do PDPH-LBU. Ela representa a indicação da largura de banda máxima permitida a um cliente durante a recepção de um vídeo. Esta banda passante é denotada por kb, onde k é o múltiplo do número de canais em relação à taxa de consumo b. No PDPH, o momento no qual a largura de banda exigida é máxima ocorre durante as primeiras m fatias de tempo, quando o cliente recupera todos os n segmentos recebidos simultaneamente. Sendo $\frac{b}{i}$ a largura de banda utilizada pelo i-ésimo canal, obtém-se, então, a seguinte restrição:

$$b\Big(H(m+n-1) - H(m-1)\Big) \leq bk$$

ou, simplificando,

$$H(m+n-1) - H(m-1) \leq k, \tag{5.9}$$

já que b deve ser estritamente positivo.

Note que o cliente precisa esperar durante as primeiras m fatias de tempo para armazenar completamente o primeiro segmento, já que é transmitido a uma largura de banda igual a $\frac{b}{m}$, e só então inicia a exibição do vídeo. Tendo cada fatia de tempo uma duração equivalente a S/n segundos, e dado um limite w para o tempo de espera máximo, tem-se a obrigatoriedade deste limite w não ser menor que o tempo de m fatias de tempo, obtendo-se então a restrição de número 2 (Inequação 5.6).

Rearranjando-se a expressão, obtém-se ainda:

$$m\frac{S}{w} \leq n, \tag{5.10}$$

um limite inferior para a variável n, que é limitada superiormente por um parâmetro n_{max} dado. A variável m é restringida através da terceira restrição, dada pela inequação 5.7.

5.5.2 PDPH-LBU com vários conjuntos de canais

Atrasos no calendário de recepção de alguns segmentos são introduzidos (Na Figura 5.5, correspondem ao sexto, sétimo e oitavo segmentos), de forma a se poder obter uma solução viável tal que o cliente receba no máximo uma quantidade de segmentos equivalente a sua disponibilidade de banda passante. Esta é a idéia do Protocolo de Difusão Poliharmônica com Limitação de Banda do Usuário (*Polyharmonic Broadcasting with User Bandwidth*

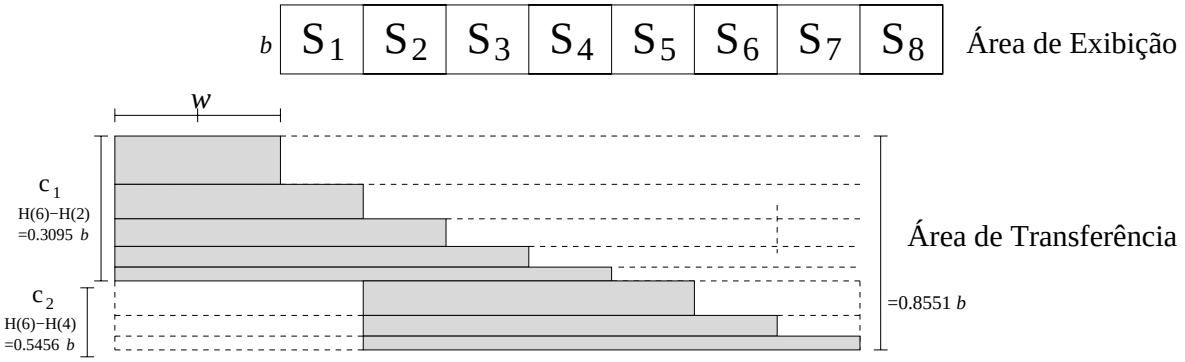

Figura 5.5: Mapa do PDPH-LBU (fora de escala em y) para $d = 2$

Limit - PDPH-LBU). A adição de conjuntos de canais acarreta em uma perda de eficiência decorrente do aumento na demanda de banda passante do servidor, mas em compensação consegue satisfazer condições mais restritivas de tempo de espera, pois assim torna-se possível o aumento do número de segmentos, diminuindo assim a latência de exibição. Há então um compromisso entre um aumento na largura de banda do servidor, e a latência de exibição, que consegue se reduzir.

Define-se um conjunto de canais como sendo um grupo de canais que têm sua recepção de segmentos iniciada simultaneamente. Na Figura 5.5, por exemplo, vê-se dois conjuntos de canais: c_1, que consiste nos cinco primeiros canais recebidos imediatamente após a requisição de exibição do cliente, e c_2, compreendendo os três últimos canais, cuja recepção só se inicia com três fatias de tempo de atraso. Durante a quarta fatia de tempo ocorre um período de transição, onde o cliente passa a receber segmentos de canais de conjuntos adjacentes, até o término da recepção de segmentos do conjunto anterior. Denota-se por d a *dimensão* do problema, em outras palavras o número de conjuntos de canais existentes em uma instância LPHB.

5.5.3 Formulação do problema para vários conjuntos de canais

A formulação do PDPH-LBU é dada por:

$$\min \sum_{i=1}^{d} \Big[H(m_i + n_i - 1) - H(m_i - 1) \Big] \tag{5.11}$$

sujeito às restrições:

$$H(m_1 + n_1 - 1) - H(m_1 - 1) \leq k \tag{5.12}$$

$$H(m_{i-1} + n_{i-1} - 1) - H(m_{i-1} + n_{i-1} - m_i) + H(m_i + n_i - 1) - H(m_i - 1) \leq k, \quad i = 2, \ldots, d \tag{5.13}$$

$$m_{i-1} + n_{i-1} - m_i \geq 0, \qquad i = 2, \ldots, d \tag{5.14}$$

$$n = \sum_{i=1}^{d} n_i \tag{5.15}$$

$$m_1 \frac{S}{w} \leq n \tag{5.16}$$

$$1 \leq m_i \leq m_{max}, \qquad i = 1, \ldots, d \tag{5.17}$$

$$1 \leq n_i \leq n_{max}, \qquad i = 1, \ldots, d \tag{5.18}$$

Neste problema de otimização, cada uma das variáveis m e n do PDPH-LBU com um conjunto de canais é substituída por d outras variáveis, cada qual relativa ao próprio conjunto de canais. Em outras palavras, cada conjunto de canais c_i possui seus próprios parâmetros m_i e n_i. Os significados são os mesmos do problema com um conjunto de canais. Assim, m_i corresponde ao termo inicial da série harmônica correspondente à largura de banda alocada ao primeiro canal de c_i, e n_i corresponde ao número de segmentos do c_i-ésimo canal.

A função objetivo do problema é obtida da mesma forma que a versão com apenas um conjunto de canais, sendo que cada conjunto de canais c_i requer uma largura de banda total igual a

$$b\Big[H(m_i + n_i - 1) - H(m_i - 1)\Big] \tag{5.19}$$

Somando a largura requerida por cada um dos conjuntos de canais e eliminando o termo multiplicativo b, obtém-se a função objetivo refletida em (5.11).

A variável n neste problema reflete agora a soma de cada valor n_i, ou seja, o número total de segmentos do vídeo, exatamente a restrição dada pela Equação 5.15.

A duas primeiras restrições (Inequações 5.12 e 5.13) são as restrições equivalentes à restrição do problema com um conjunto de canais (5.5) e diz respeito à largura de banda máxima permitida para difusão. A diferença é que há uma restrição somente para o primeiro conjunto de canais (Inequação 5.12), e na segunda restrição (Inequação 5.13) há $(d-1)$ restrições para cada fase de transição entre dois conjuntos de canais adjacentes.

A restrição seguinte (Inequação 5.14) torna obrigatório que um conjunto de canais não inicie a recepção antes que o conjunto anterior. O último segmento do conjunto de canais anterior $(i-1)$ é transmitido em $m_{i-1}+n_{i-1}-1$ fatias de tempo. O primeiro segmento do conjunto de canais atual i, por sua vez, é transmitido em m_i fatias de tempo. A diferença entre estes dois valores não pode ser superior a uma fatia de tempo, que corresponde ao intervalo de tempo entre o fim da transmissão de um canal para outro adjacente. Em outras palavras, como se trata da transição de um conjunto de canais para outro, neste caso ocorre uma aglutinação dos conjuntos de canais.

A restrição de latência de exibição do PDPH-LBU com vários conjuntos de canais (Inequação 5.16) é obtida da mesma maneira que no modelo com apenas um conjunto (vide Inequação 5.6), observando que o cliente deve esperar somente durante as primeiras m_1 fatias de tempo, e que uma fatia de tempo equivale ao tempo total do vídeo, S, dividido pelo número total de segmentos, ou seja, $\sum_{i=1}^{d} n_d$. Tem-se então:

$$w \geq m_1 \frac{S}{n_1 + n_2 + \cdots + n_d},$$

ou, rearranjando:

$$n_1 + n_2 + \cdots + n_d \geq m_1 \frac{S}{w} \tag{5.20}$$

Substituindo através da Equação 5.15, obtém-se então a restrição dada pela Inequação (5.16).

A restrição dada pelas Inequações (5.17) e (5.18) tem a finalidade de limitar o valor de cada uma das variáveis m_i e n_i, onde $1 \leq i \leq d$.

5.5.4 Mapeamento do Problema para Algoritmos Genéticos

Os problemas do PDPH-LBU são problemas de otimização inteiros e não-lineares. Não existe algoritmo computacional para a solução específica destes problemas. Assim sendo, o problema de PDPH-LBU foi resolvido, para $d = 1$ e $d = 2$, via avaliação de todas as

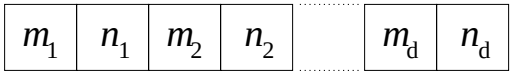

Figura 5.6: Estrutura de um cromossomo para resolução de uma instância PDPH-LBU

possíveis soluções, e para $d \geq 2$, foram utilizados também algoritmos genéticos, baseados a partir da referência [28].

A estrutura de um cromossomo para uma solução PDPH-LBU (Figura 5.6) possui relação direta com as variáveis envolvidas no problema de otimização. Como cada solução depende apenas dos parâmetros m e n de cada conjunto de canais, cada cromossomo é composto justamente pelos pares de parâmetros m_i e n_i, para cada conjunto de canais $c_1, c_2, \ldots, c_d$.

A população inicial é gerada de forma que para cada gene de um indivíduo receba um valor aleatório entre sua faixa de valores possíveis. Depois da geração dos genes, são checadas as restrições do problema, para verificação da factibilidade da solução. Caso contrário, o indivíduo é descartado e não entra na população. Desta forma, os indivíduos são adicionados até atingir o tamanho da população inicial, definido neste trabalho em 500 indivíduos. Se ocorrerem 100.000 tentativas improdutivas para gerar um indivíduo, o melhor indivíduo é apresentado como solução do problema. A mesma verificação das restrições realizada na fase de geração dos indivíduos é feita após a geração de novos indivíduos por cruzamento e mutação (cujas taxas, iguais a 60% e 5%, respectivamente, sã indicadas em [27]). Neste trabalho, utiliza-se 100 gerações para a obtenção do resultado final.

A fase de seleção dos indivíduos foi feita utilizando-se uma roleta com função de aptidão (*fitness*) igual a $\frac{1}{(F_{obj})^2}$. Também foi utilizada a reprodução seletiva, que garante que a melhor solução gerada até então sobreviva à próxima geração. O espaço de busca para a otimização da função de PDPH-LBU torna-se muito grande, o que faz com que uma heurística que não faça avaliação de todos as possíveis soluções, como algoritmos genéticos, obtenha uma solução aproximada da solução ótima, mas com um tempo de processamento muito menor.

5.5.5 Uma avaliação da efetividade do PDPH-LBU

Metodologia utilizada

Para se avaliar a efetividade do protocolo PDPH-LBU, foram realizados estudos sobre o compromisso entre as limitações do lado do cliente, notadamente a largura de banda, o tempo de espera máximo ao qual o cliente é submetido, e a largura de banda necessária pelo servidor no protocolo PDPH-LBU. Para tal fim, exemplos numéricos foram derivados.

Os problemas de otimização foram desenvolvidos em C e executados em ambiente Linux em PCs Pentium 4.

O tempo de execução para a obtenção de uma solução final no PDPH-LBU depende da abordagem utilizada. No caso da avaliação de todas as soluções possíveis, o tempo depende diretamente dos parâmetros m_{max} e n_{max}. Para $d = 1$ o tempo de execução não foi significativo, assim como também não o foi para $d = 2$ com valores de m_{max} e n_{max} até 100. Já para $m_{max} = n_{max} = 1000$, cada solução final chega a levar uma hora de processamento. Para $d = 3$ o tempo de processamento tornou-se inviável, devido ao crescimento exponencial característico do problema. Utilizando-se algoritmos genéticos, soluções para $d = 1$ e $d = 2$ foram encontradas em tempo não significativo, e para $d = 3$, demora-se entre cinco e dez minutos para gerar cada solução final. A variação deste tempo se deve em grande parte à fase inicial do algoritmo, durante a geração da população inicial, pois a medida em que as soluções tornam-se mais restritivas, mais difícil torna-se também a geração — aleatória — de um indivíduo que satisfaça estas condições. Quando possível, os dados gerados através de algoritmos genéticos foram confrontados com a avaliação de todas as soluções possíveis e não apresentaram diferenças significativas, citando-se como exemplo o caso com dois conjuntos de canais e uma restrição $k = 5$. Para estes parâmetros, obteve-se a mesma latência de espera (0,6% do tempo de duração total do vídeo), tendo a solução por algoritmos genéticos requerido $0,02b$ de banda passante a mais que avaliação de todas as soluções possíveis ($5,71b$ contra $5,69b$).

Resultados Numéricos

Para os gráficos a seguir, utiliza-se a fração w/S como eixo das abcissas, uma vez que todos os resultados de latência de exibição obtidos podem ser dados em função do tempo total do vídeo. Exemplificando, um resultado onde $w/S = 0,04$ equivale a dizer que o tempo de espera para exibição é de 4% do tempo de duração total. Para um clipe musical de cinco minutos, esta latência equivale a doze segundos, enquanto em um filme de duas horas corresponde a quase cinco minutos de espera. Os resultados nas ordenadas estão

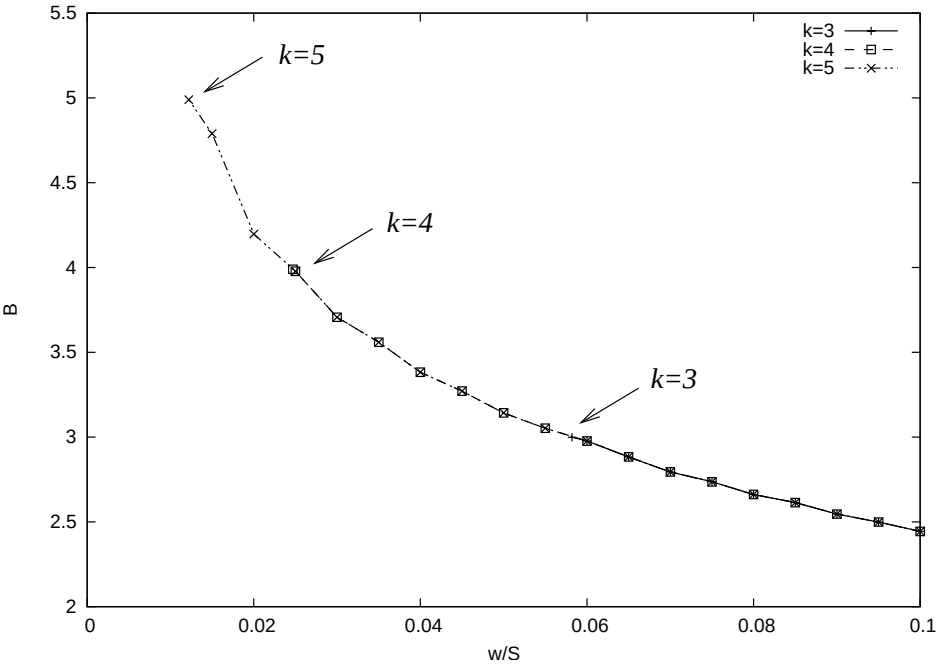

Figura 5.7: Valores ótimos de B em função de w/S para PDPH-LBU ($d = 1$)

em função da largura de banda do servidor (função B), e é dada em múltiplos da taxa b utilizada para transmissão de um fluxo de vídeo, geralmente dada em Mbps. Uma solução encontrada que utilize uma largura de banda igual a $3,5b$ pode gerar uma demanda no servidor de 392 kbps se se utiliza um fluxo MPEG-4 de 112 kbps, ou 70 Mbps em caso de utilização de um fluxo MPEG-2 HDTV de 20 Mbps.

Na Figura 5.7, mostra-se a demanda por largura de servidor (dada pela função B), em função da razão entre a latência de exibição e o tamanho total do vídeo, w/S. O gráfico mostra o limite de tempo de espera de acordo com o número de canais que o cliente possui. No caso do cliente possuir três canais ($k = 3$), a fração w/S equivale a 0,0582. Isto significa que, para um filme de duas horas de duração este cliente deverá esperar até oito minutos. Para $k = 4$ e $k = 5$, o cliente deve esperar, respectivamente, três e um minuto. Pode-se observar ainda a sobreposição de todas as curvas, independentemente do número de canais do cliente. A diferença entre elas está no limite inferior w/S obtido. Para cada curva, são geradas várias soluções para valores de w/S diferentes, até o ponto onde não se encontram soluções factíveis, sendo este ponto considerado como o limite inferior de w/S.

Até o ponto onde $w/S = 0,0582$, limite da latência para $k = 3$, as três curvas permanecem sobrepostas. Valores menores de latência requerem uma banda passante do

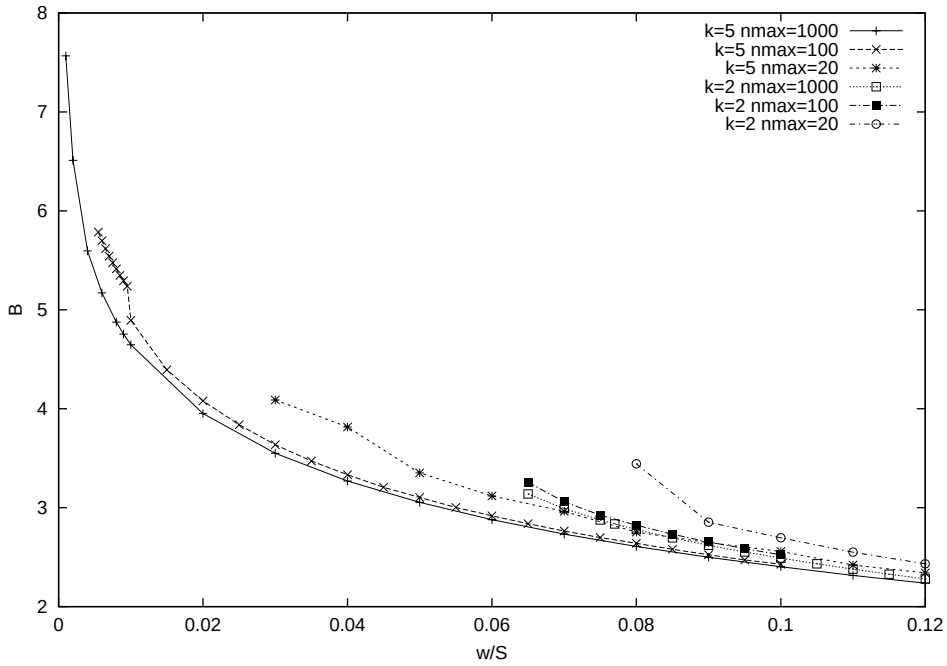

Figura 5.8: Estudo comparativo de n_{max} para o PDPH-LBU ($d = 2$)

cliente e do servidor maior do que $3b$ (uma vez que só há um conjunto de canais e todos devem ser recuperados simultaneamente), o que inviabiliza estas soluções. Entre as faixas de w/S entre 0,0582 e 0,0247, as curvas $k = 5$ e $k = 4$ se sobrepõem. Novamente, para latências menores do que 2,47% do vídeo são requeridos mais do que $4b$, inviabilizando assim as soluçõs para $k = 4$. A curva $k = 5$ persiste até a latência de 0,0122, ou um minuto e 12 segundos para um vídeo de duas horas.

Como este gráfico ilustra o PDPH-LBU com apenas um conjunto de canais, toda a largura de banda disponível pelo cliente é utilizada pelo servidor. Por este motivo, as curvas acabam sobrepostas, uma vez que a única limitação que as diferencia está na própria restrição de largura de banda máxima do cliente (restrição 5.5 do PDPH-LBU). Por exemplo, ao desejar um tempo de espera igual a 4% do tamanho do vídeo, serão necessários $3.38b$ de largura de banda, permitidos apenas para clientes com largura de banda igual ou superior a esta. Estes valores evidenciam o compromisso entre a largura de banda dos clientes e a latência de exibição.

A Figura 5.8 mostra a influência da escolha de n_{max} para a resolução de PDPH-LBU para valores de k iguais a dois e cinco. As curvas do gráfico mostram um aumento significativo na demanda por banda passante e tempo de espera para uma segmentação pequena ($n = 20$) em relação a uma segmentação grande ($n = 1000$). Uma alternativa

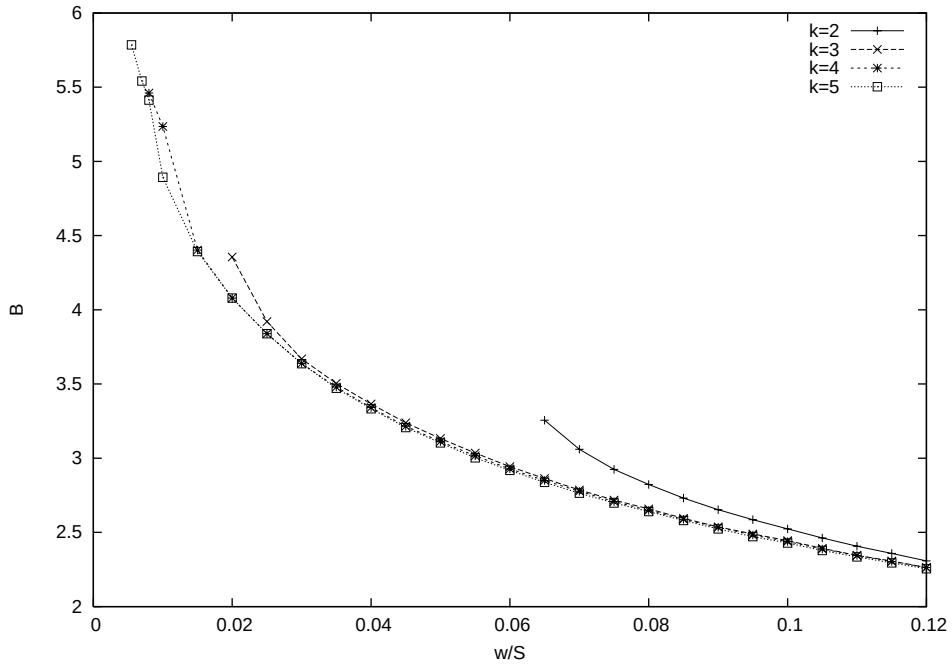

Figura 5.9: Largura de banda do servidor em função de w/S para PDPH-LBU ($d = 2$)

pode ser a utilização de uma segmentação intermediária que não requeira muita largura de banda adicional e ao mesmo tempo possua bons resultados em relação ao tempo de espera, como no caso de $n = 100$. Cabe salientar que uma segmentação excessiva poderia dificultar o processamento em servidores, acarretando um aumento de custos na infra-estrutura.

A segmentação exerce um papel direto na latência de exibição, porque um número maior de segmentos acarreta na diminuição do tamanho geral dos segmentos. Como a latência é diretamente relacionada com o tamanho do primeiro segmento, esta é reduzida na mesma medida.

Percebe-se ainda no gráfico que para uma restrição de $k = 2$ ocorre uma queda abrupta de desempenho em relação à restrição $k = 5$; este resultado já era esperado, uma vez que na referência [10] são indicadas perdas significativas quando há restrição no cliente de utilização de até dois canais simultâneos. Para uma segmentação de até 20 segmentos por conjunto de canais, a melhor latência obtida é de 8%, o que equivale a aproximadamente 8 minutos e meio para um vídeo de duas horas. Mesmo maiores segmentações não conseguem reduzir substancialmente a latência, baixando para sete minutos e 48 segundos.

Pode se constatar na Figura 5.9 uma pequena diferença entre as curvas, diferentemente da sobreposição encontrada na Figura 5.7. Vê-se nesta diferença uma pequena melhoria

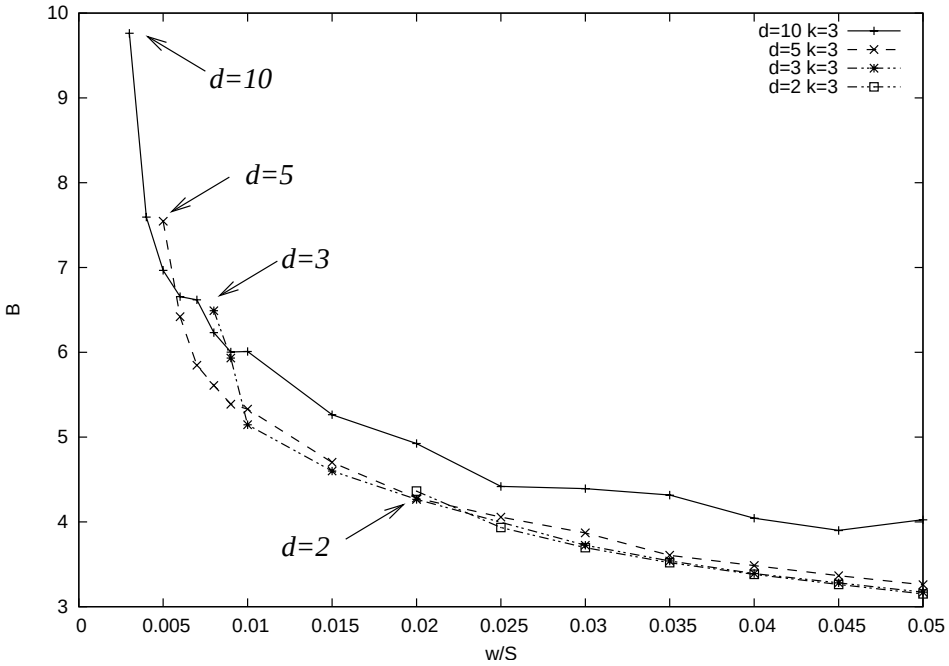

Figura 5.10: Influência do parâmetro d

no requerimento da largura de banda do servidor quando o cliente possui uma maior capacidade de largura de banda. Nota-se também uma nítida vantagem em relação ao tempo de espera quando os clientes possuem uma largura de banda maior. Tanto a vantagem nos valores de B quanto na obtenção de melhores latências devem-se ao fato da abertura de novas possibilidades de configuração abertas pelo uso de dois conjuntos de canais. Para restrições mais apertadas, como $k = 2$ acabam sem poder beneficiar-se de tais configurações.

Clientes com limitações de largura de banda de $k = 4$ e de $k = 5$, isto é, com largura de banda 4 e 5 vezes maior que a taxa de exibição do vídeo podem ser atendidos mesmo quando o tempo total do vídeo é superior a 125 e 180 vezes o tempo de espera, respectivamente. Ou seja, em torno de 58 (para $k = 4$) e 40 segundos (para $k = 5$) de espera para um vídeo de duas horas. Em comparação com a versão com um conjunto de canais (Figura 5.7), percebe-se a diminuição da latência de exibição em todos os casos. Na curva com a limitação $k = 5$, consegue-se um tempo de espera equivalente a 0,0055 do vídeo todo, enquanto que para $d = 1$ o valor mínimo obtido é de 0,0122.

A Figura 5.10 ilustra a influência do parâmetro d no protocolo PDPH-LBU. Percebe-se uma clara melhoria devido à grande redução na latência mínima obtida com o aumento do número de conjuntos de canais. Da latência mínima de 5,82% do tamanho do vídeo

para um conjunto de canais (Figura 5.7) para 2% com dois conjuntos de canais, chegando a 0,8% para $d = 3$. A latência ainda é reduzida para 0,5% quando $d = 5$ e $0,3\%$ para $d = 10$. Para o caso com cinco conjunto de canais, percebe-se uma demanda ligeiramente maior do que para os casos com dois e três conjuntos de canais. Para $d = 10$, ao contrário dos anteriores, nota-se o grande aumento de demanda na largura de banda do servidor (em média, $0,8b$ a mais). No PDPH-LBU, a adição de um conjunto de canais implica no aumento de até n_{max} segmentos extras, o que diminui drasticamente a latência de exibição na transição de um até três conjuntos de canais. Desta forma, obtém-se soluções com pouca ou nenhuma diferença na demanda de banda passante do servidor, antes impossíveis por não atingir a latência especificada. Já para dimensões maiores o ganho na latência não é significativo (como na Figura 5.8, onde o desempenho com 1000 segmentos não é significativamente superior às soluções com 100 segmentos), aumentando assim de forma gradativa a largura de banda necessária.

5.6 GEBB-LBU

O protocolo GEBB também é um protocolo otimamente estruturado (como visto na Seção 4.17). Da mesma forma que o protocolo PDPH-LBU (Seção 5.5), esta seção aborda o desenvolvimento de soluções para o GEBB mediante limitação na banda passante de usuários. O novo protocolo introduzido aqui denomina-se Protocolo de Difusão Guloso com Canais de Mesma Largura de Banda com Limitação de Banda do Usuário (*Greedy Equal Bandwidth Broadcasting with Limited User Bandwidth*), ou GEBB-LBU.

O desenvolvimento do protocolo segue exatamente os mesmos passos do PDPH-LBU: constrói-se o protocolo com apenas um conjunto de canais e por fim generaliza-se o modelo através da adição de novos conjuntos de canais.

É importante ressaltar uma diferença na característica entre os protocolos PDPH-LBU e GEBB-LBU. No PDPH sem restrições o modo de segmentação é conhecido, sendo a divisão feita em tamanhos iguais. Já no GEBB sem restrições, busca-se uma segmentação ótima baseada nos princípios dos algoritmos otimamente estruturados. O problema de otimização do GEBB-LBU deve determinar então um particionamento de vídeo — incluindo-se aí a largura de banda para cada canal correspondente a um segmento — tal que requeira a menor banda passante possível para o servidor.

5.6.1 GEBB-LBU com um conjunto de canais

Quando existe a limitação da largura de banda dos clientes, pode-se utilizar as equações do GEBB (que originalmente partem de S e w arbitrários para encontrar a largura de banda necessária) para encontrar qual o tempo de espera necessário para uma transmissão de um vídeo que utilize k canais de largura de banda b.

Obtém-se primeiramente a largura de banda correspondente a cada um dos canais em função dos parâmetros k e n, que são arbitrados como dados do problema. Normalizando a largura de banda de consumo (isto é, fazendo $b = 1$), temos:

$$b_i = \frac{k}{n} \tag{5.21}$$

Substituindo em (4.15) e rearranjando, obtém-se:

$$w = \frac{S}{\left(\frac{k}{n} + 1\right)^n - 1} \tag{5.22}$$

Ao normalizar também o tamanho total do vídeo ($S = 1$), tem-se:

$$w = \frac{1}{\left(\frac{k}{n} + 1\right)^n - 1} \tag{5.23}$$

Os tamanhos dos segmentos S_i são obtidos da mesma maneira que no GEBB, ou seja,

$$S_i = wb_i(1 + b_i)^{i-1} \tag{5.24}$$

Da mesma forma que no protocolo PDPH-LBU, o parâmetro n é dado em GEBB-LBU para limitar a complexidade de gerenciamento dos segmentos, tanto na *set-top-box* quanto no servidor.

5.6.2 GEBB-LBU com vários conjuntos de canais

Da mesma forma que no protocolo PDPH-LBU com apenas um conjunto de canais, nem sempre é possível encontrar soluções viáveis para parâmetros fornecidos. Portanto, procura-se agora estender o protocolo com um raciocínio de construção semelhante ao do PDPH-LBU: dividem-se os canais em conjuntos, onde um conjunto de canais pode ter um atraso no calendário de recepção do cliente em relação ao conjunto anterior, para que o cliente que possua limitação de largura de banda possa receber uma quantidade máxima possível de dados (Figura 5.11).

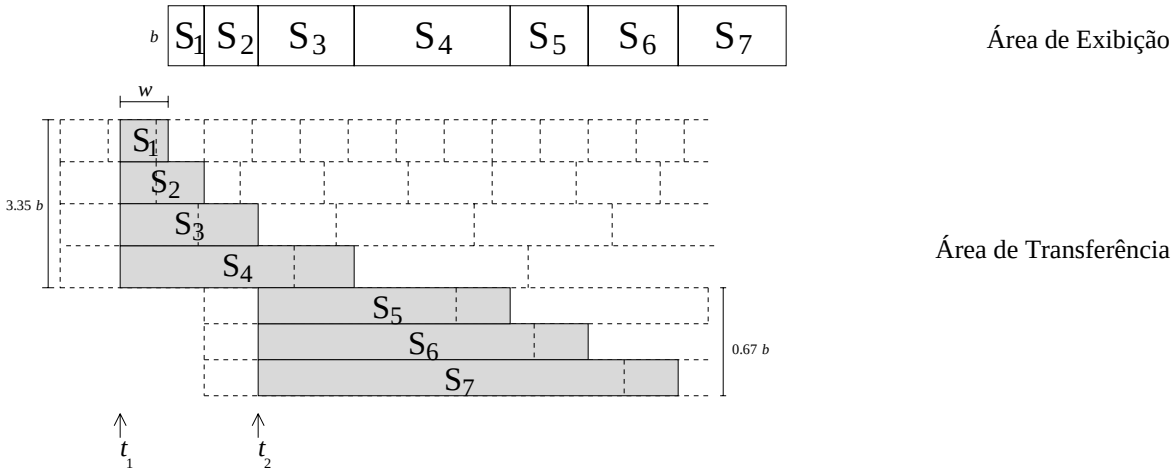

Figura 5.11: Mapa de Difusão do GEBB-LBU ($d = 2$)

A extensão do PDPH-LBU permitiu a utilização de variáveis somente de conjuntos de canais. No GEBB-LBU, isto não é possível, pois neste caso a segmentação desejada não é conhecida, sendo parte da solução. Por este motivo, utiliza-se dois tipos de variáveis. O primeiro tipo engloba as que se aplicam a um único canal ou segmento, como no caso das variáveis S_i e b_i ($i = 1, \ldots, n$), que indicam o tamanho do i-ésimo segmento e a largura de banda alocada para sua transmissão, respectivamente. Já o segundo tipo compreende as variáveis que se aplicam aos conjuntos de canais. É o caso das variáveis $s^{(c)}$, $n^{(c)}$ e $a^{(c)}$, onde c estende-se de 1 a d). Estas últimas variáveis possuem uma notação diferenciada, cujo índice aparece sobrescrito e entre parênteses, para facilitar na diferenciação das mesmas, uma vez que são de "naturezas" diferentes.

Para cada conjunto de canais c ($c = 1, \ldots, d$), existem três variáveis denominadas $s^{(c)}, n^{(c)}$ e $a^{(c)}$. As duas primeiras dizem respeito, respectivamente, à soma do tamanho dos segmentos que compõem o c-ésimo conjunto de canais, e ao número de segmentos acumulados até o referido conjunto de canais. Na Figura 5.11, por exemplo, tem-se o primeiro conjunto de canais, formado pelos quatro primeiros (isto é, $n^{(1)} = 4$). Já para o segundo conjunto, composto por três canais, tem-se $n^{(2)} = 7$. A decisão de se utilizar o número acumulado de segmentos contidos no conjunto de canais é feita de forma a facilitar tanto na modelagem do problema quanto a implementação da solução através de algoritmos genéticos. A definição da variável $a^{(c)}$ é relacionada ao momento de término de recepção do último segmento anterior ao conjunto de canais c. O seu significado corresponde ao número de segmentos com atraso de recepção em relação ao primeiro conjunto de canais.

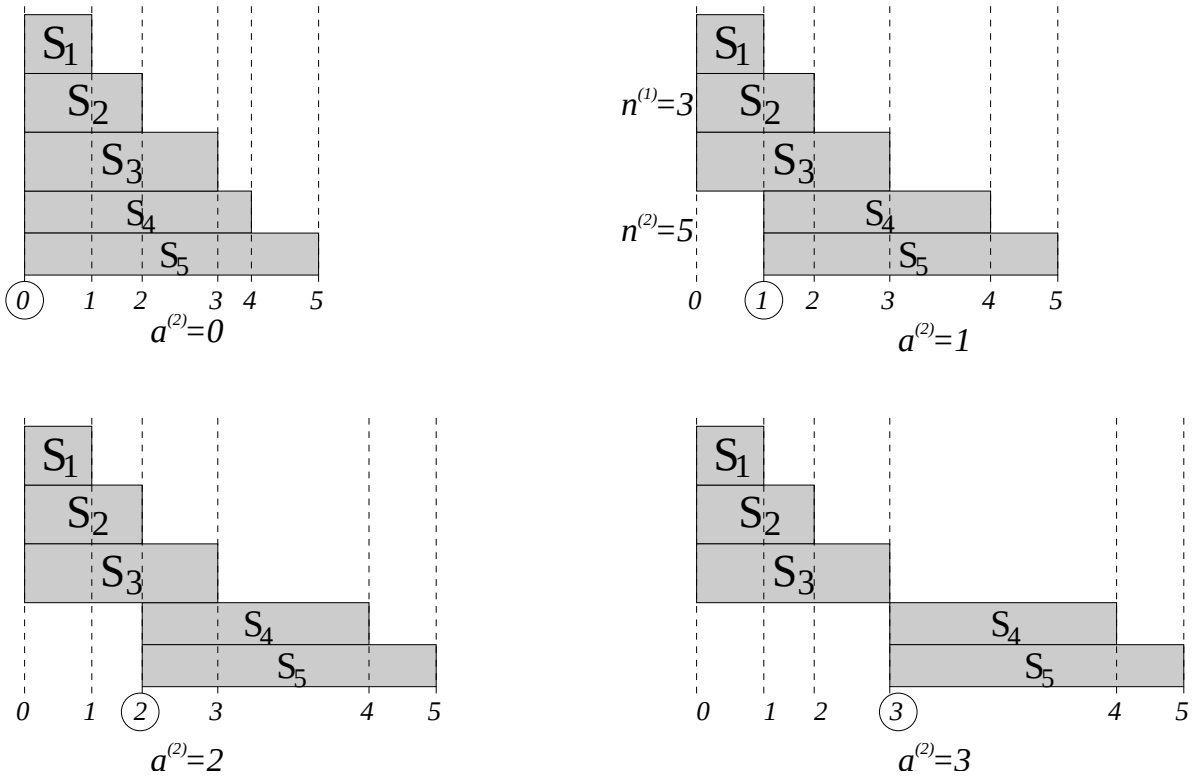

Figura 5.12: Possíveis valores para a variável $a^{(2)}$ ($n^{(1)} = 3$, $n^{(2)} = 5$)

Na Figura 5.12, são ilustrados os valores possíveis de $a^{(2)}$ para uma configuração de cinco canais no total, sendo $n^{(1)} = 3$ e $n^{(2)} = 2$. Cada variável $a^{(c)}$, em outras palavras, deve estar entre zero (Quando então o protocolo funciona como se o conjunto c fundisse com o conjunto anterior, ocorrendo o mesmo sempre que $a^{(c)} = a^{(c-1)}$) e n_i. Por convenção, assume-se que $a^{(1)}$ deva sempre ser igual a zero, uma vez que não existem conjuntos de canais anteriores ao primeiro.

O problema de otimização ligado ao GEBB-LBU é dado a seguir:

$$\min \sum_{i=1}^{n} b_i \tag{5.25}$$

sujeito às restrições:

$$b_i \left(w + \sum_{j=1}^{i-1} S_j \right) = S_i \qquad i = 1, 2, \ldots, n^{(1)} \tag{5.26}$$

$$b_i\left(w+\sum_{j=1}^{i-1}S_j\right)=S_i \qquad i=n^{(c-1)}+1,n^{(c-1)}+2,\ldots,n^{(c)} \text{ se } a^{(c)}=0, \qquad \forall c\in 2\ldots d \tag{5.27}$$

$$b_i\left(\sum_{j=a^{(c)}}^{i-1}S_j\right)=S_i \qquad i=n^{(c-1)}+1,n^{(c-1)}+2,\ldots,n^{(c)} \text{ se } a^{(c)}>0, \qquad \forall c\in 2\ldots d \tag{5.28}$$

$$\sum_{j=a^{(c)}+1}^{n^{(c)}}b_j\leq k \qquad c=1,\ldots,d \tag{5.29}$$

$$\begin{array}{ccccccc} b_1 & = & b_2 & = & \ldots & = & b_{n^{(1)}} \\ b_{n^{(1)}+1} & = & b_{n^{(1)}+2} & = & \ldots & = & b_{n^{(2)}} \\ \vdots & & \vdots & & \ldots & & \vdots \\ b_{n^{(d-1)}+1} & = & b_{n^{(d-1)}+2} & = & \ldots & = & b_{n^{(d)}} \end{array} \tag{5.30}$$

$$\sum_{i=1}^{n^{(1)}}S_i=s^{(1)} \tag{5.31}$$

$$\sum_{i=n^{(c-1)}+1}^{n^{(c)}}S_i=s^{(c)}, \qquad c=2,\ldots,d \tag{5.32}$$

$$\sum_{c=1}^{d}s^{(c)}=S \tag{5.33}$$

$$0=a^{(1)}\leq a^{(2)}\leq\cdots\leq a^{(d)} \tag{5.34}$$

$$a^{(c)}\leq n^{(c-1)} \qquad c=2,\ldots,n \tag{5.35}$$

$$
\begin{aligned}
& w > 0 \\
& 0 < S_i < S, \qquad i = 1, 2, \ldots, n \\
& 0 < b_i, \qquad\quad i = 1, 2, \ldots, n \\
& 1 < n \\
& 1 < d
\end{aligned}
\tag{5.36}
$$

A adição de c conjuntos de canais ao conjunto original e as variáveis criadas correspondentes a cada conjunto gera várias restrições para relacionar estas variáveis com as variáveis que quantificam os canais e segmentos (respectivamente, b_i e S_i).

Para os canais do primeiro conjunto $c^{(1)}$, utiliza-se a Equação 5.26 da primeira restrição, para fazer a correspondência entre um segmento e os anteriores. Esta restrição é idêntica à Equação 4.10 do problema GEBB, excetuando-se seu escopo.

Na segunda restrição (Equações 5.27 e 5.28), são considerados os aspectos de continuidade de exibição para os segmentos pertencentes aos conjuntos restantes. Para estes segmentos utiliza-se a Equação (equação 5.27) quando $a^{(c)}$ for igual a zero, ou a Equação (equação 5.28) quando for maior que zero. Para efeito de facilidade na implementação do algoritmo, cria-se uma variável adicional S_0, cujo valor é equivalente à latência de exibição w. Desta forma, utiliza-se apenas a Equação 5.28.

A terceira restrição (Equação 5.29) compreende na verdade d restrições, uma para cada conjunto de canais. Cada uma delas limita a largura de banda máxima no instante no qual um conjunto de canais tem sua recepção iniciada, quando o requisito de banda passante é um possível máximo.

As equações contidas em 5.30 refletem a relação existente entre as variáveis individuais e as variáveis de conjuntos de canais. Especificamente, a largura de banda é a mesma para todos os canais pertencentes a um conjunto.

Da mesma forma ocorre com as restrições no tamanho dos segmentos, que seguem duas regras. A primeira delas diz que a soma dos tamanhos dos segmentos de um conjunto equivale ao tamanho do conjunto (Equações 5.31 e 5.32). Na segunda, o total do tamanho dos conjuntos deve ser igual ao tamanho total do vídeo.

A restrição 5.35 faz com que todo conjunto de canais tenha um atraso em relação ao conjunto anterior (quando $a^{(c)} < a^{(c-1)}$). No caso de $a^{(c)} = a^{(c-1)}$ para algum conjunto c, o conjunto é aglutinado com o anterior. Quando $a^{(1)} = a^{(2)} = \cdots = a^{(d)} = 0$, a formulação é idêntica à do GEBB-LBU com somente um conjunto de canais.

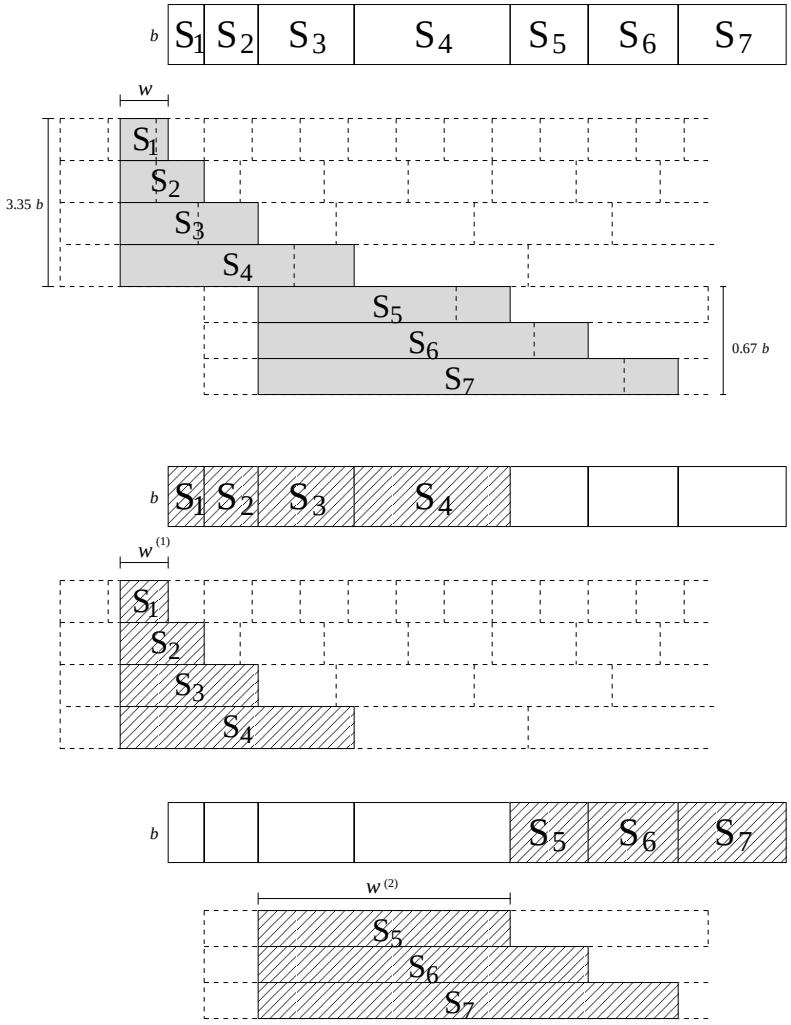

Figura 5.13: Um problema do GEBB-LBU tratado como d problemas do GEBB

Cabe aqui salientar uma diferença na restrição de número de segmentos dos problemas de otimização do PDPH-LBU e do GEBB-LBU. O PDPH-LBU utiliza um máximo de n_{max} segmentos por conjunto de segmentos. Isto não acontece com o GEBB-LBU, que possui uma restrição mais apertada, uma vez que o número máximo de segmentos n é distribuído ao longo dos conjuntos de canais.

Cálculo do tamanho dos segmentos e da largura de banda dos canais

A forma como foi estruturado o problema de otimização pode trazer benefícios em sua implantação, tratando cada conjunto de canais como se fosse um problema GEBB isolado. A Figura 5.13 mostra em sua parte superior um mapa de difusão do GEBB-LBU com

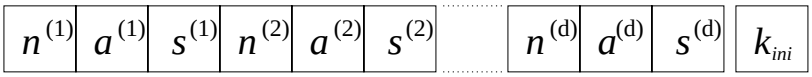

Figura 5.14: Estrutura de um cromossomo para resolução de uma instância GEBB-LBU

$d = 2$, juntamente com os dois mapas GEBB equivalentes ao problema: na parte central e inferior da figura, respectivamente, os mapas para o primeiro conjunto de canais (que compreende os canais um a quatro), e o segundo conjunto de canais (cinco a sete).

Esta característica dos conjuntos de canais permite a utilização das equações do GEBB sem restrições (Equações 4.16 e 4.17), fazendo-se a mudança apropriada de variáveis. No primeiro conjunto de canais, por exemplo, o tamanho do vídeo corresponde apenas aos segmentos que vão de S_1 a S_4. Já no segundo conjunto, de S_5 a S_7. Ainda neste conjunto, deve-se ajustar a variável w, que para o conjunto corresponde ao tempo de exibição dos segmentos S_3 e S_4, como indicado na Figura 5.13.

Mapeamento do Problema para Algoritmos Genéticos

Da mesma forma que no protocolo PDPH-LBU, a introdução de variáveis inteiras torna a solução algébrica do GEBB-LBU em uma tarefa complexa. Por essa razão, o GEBB-LBU também foi modelado utilizando-se algoritmos genéticos.

A estrutura de um cromossomo para uma solução GEBB-LBU (Figura 5.14) é composta pelos parâmetros que definem cada conjunto de canais. As variáveis $a^{(c)}$ são as mesmas utilizadas no problema de otimização, ou seja, representam valores inteiros que refletem o atraso (medido em número de segmentos) em relação ao primeiro conjunto de canais (Figura 5.12). Cada variável $n^{(c)}$, por sua vez, indica o número de canais que contém o conjunto de canais c, de forma acumulada. Já as variáveis $s^{(c)}$ refletem a soma dos segmentos $S_{n^{(d-1)}+1}$ até $S_{n^{(d)}}$, ou seja, a fração de tempo que os segmentos do conjunto c ocupa em relação ao tempo total do vídeo. Por fim, o gene k_{ini} define a porção de largura de banda inicial a ser utilizada pelo primeiro conjunto de canais.

Os parâmetros utilizados para os algoritmos genéticos foram basicamente os mesmos do problema PDPH-LBU: população de 500 indivíduos, com taxas de cruzamento (*crossover*) e mutação iguais a 60% e 5%, respectivamente, durante 100 gerações. Como operadores genéticos, utilizou-se cruzamento em um ponto e mutação dos cromossomos. A fase de seleção dos indivíduos foi proporcional à função de aptidão (igual a $\frac{1}{R}$ para a função objetivo, através do método da roleta. Também foi utilizada a reprodução seletiva.

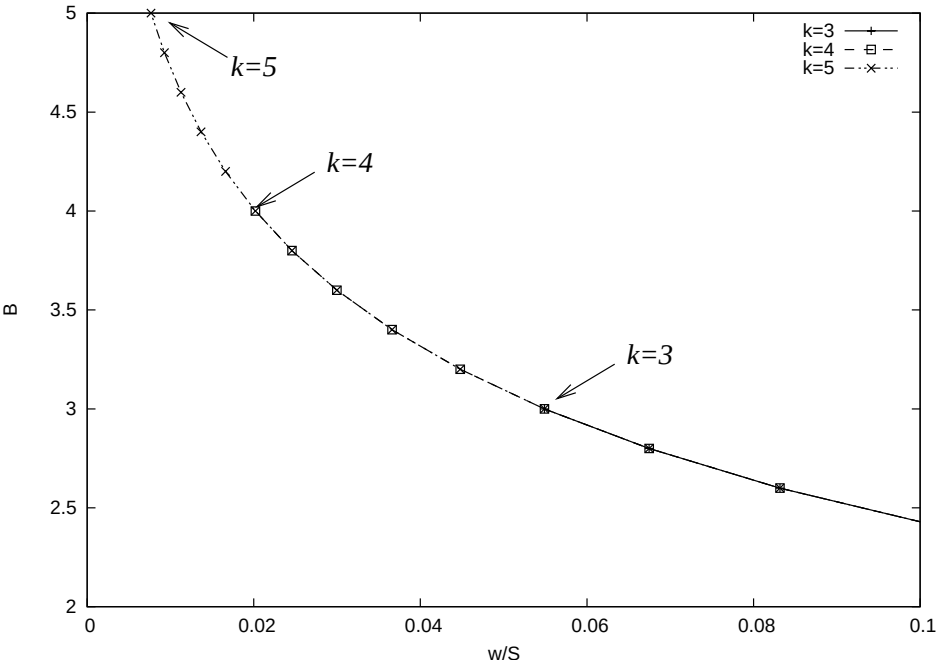

Figura 5.15: Protocolo GEBB-LBU com um conjunto de canais ($d = 1$)

5.6.3 Uma avaliação da efetividade do GEBB-LBU

O gráfico ilustrado pela Figura 5.15 indica a largura de banda demandada pelo protocolo sem conjuntos adicionais, de acordo com várias latências de exibição (como fração do tempo total do vídeo, ou seja, w/S). Da mesma forma que a Figura 5.7, pode-se notar a mesma sobreposição das curvas, e diferentes limites de latência obtidos: 5,50%, 2,02% e 0,77%, para valores de k iguais a três, quatro e cinco, respectivamente, indicados pelas setas no gráfico. Para exemplificar, em um vídeo de duas horas os tempos de espera seriam aproximadamente de seis minutos e meio, dois minutos e meio e 55 segundos, para $k = 3$, $k = 4$ e $k = 5$. Esta sobreposição de curvas ocorrerá em qualquer adaptação de um protocolo otimamente estruturado sem atrasos em canais, pois como só há um conjunto de canais, o cliente sempre tentará recuperar todos os segmentos simultaneamente, utilizando toda sua largura de banda disponível. Clientes com restrições mais apertadas como $k = 2$ não recuperam uma quantidade grande de dados, obtendo-se uma latência mínima de 16% do tempo da duração de um vídeo.

A Figura 5.16 mostra a influência do parâmetro n sobre a latência de espera para uma limitação de largura de banda a cinco canais ($k = 5$), ainda considerando-se apenas um conjunto de canais. Um aspecto interessante é o ganho de desempenho devido à diminuição da demanda de largura de banda do servidor, advindo somente através do

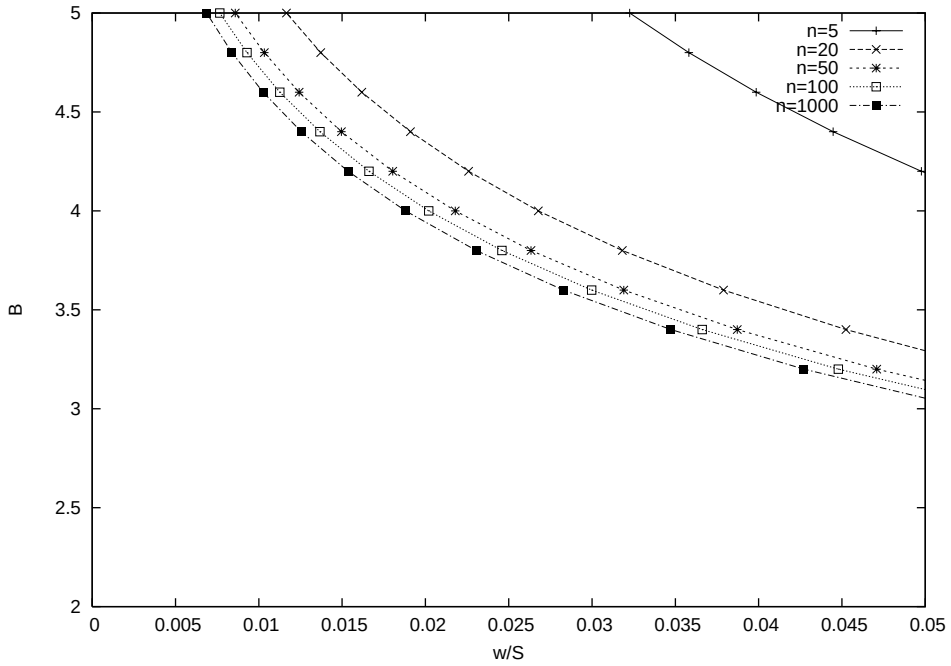

Figura 5.16: Influência do parâmetro n no GEBB-LBU ($d = 1$)

aumento da segmentação. Entretanto, este ganho diminui conforme o aumento de n. O ganho de $n = 5$ para $n = 20$ é muito maior, tanto em termos de latência de exibição quanto de banda passante, do que de $n = 20$ para $n = 100$. O mesmo ocorre para segmentações iguais a 50, 100 ou 1000, seguindo um comportamento logarítmico: a latência obtida com $k = 5$ com uma segmentação $n = 5$ é de 0,03226, e melhora drasticamente se a segmentação aumenta para $n = 20$: 0,0166, num ganho de 0,0206 em relação ao tamanho do vídeo, ou seja, o tempo de espera diminui e dois minutos e meio para um vídeo de duas horas. Por outro lado, para uma segmentação $n = 100$ consegue-se uma latência igual a 0,00766. Multiplicando por dez a segmentação, a latência obtida passa para 0,00687, traduzindo num ganho de apenas 0,00079, que traduz-se num ganho de latência de seis segundos para um vídeo de duas horas.

As Figuras 5.17 a 5.20 mostram a influência do parâmetro n (cujos valores incluem 5, 20, 50 e 100 segmentos, respectivamente) sobre o tempo de espera normalizado, cada um comparando as curvas com as restrições $k = 3$, $k = 4$ e $k = 5$. Para $k = 2$, nenhuma segmentação testada conseguiu obter uma latência de exibição. A Figura 5.20 exibe os casos para $n = 5$. Observa-se que a segmentação muito pequena traduz-se no fraco desempenho em relação às curvas correspondentes com maior segmentação. Como exemplo, pode-se citar a obtenção da latência em 0,0061 para $n = 5$, em contraste com

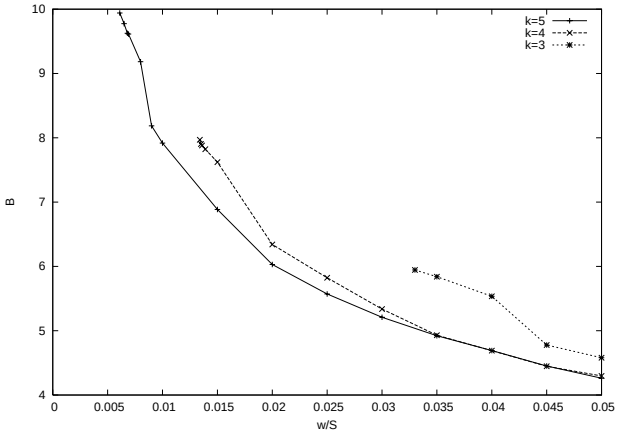

Figura 5.17: GEBB-LBU (d=2,n=5)

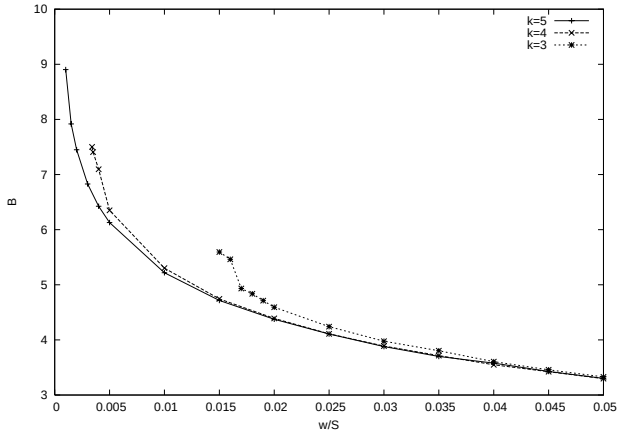

Figura 5.18: GEBB-LBU (d=2,n=20)

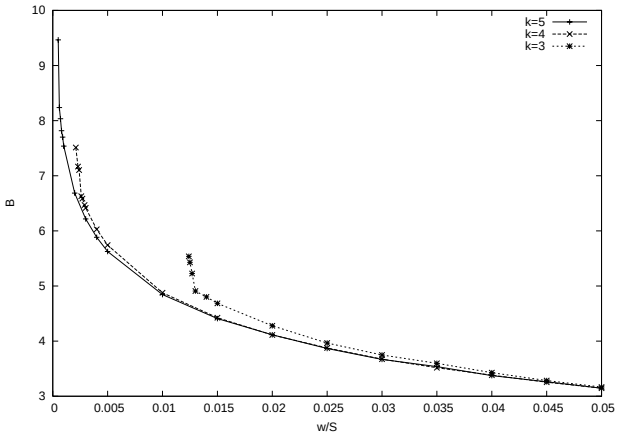

Figura 5.19: GEBB-LBU (d=2,n=50)

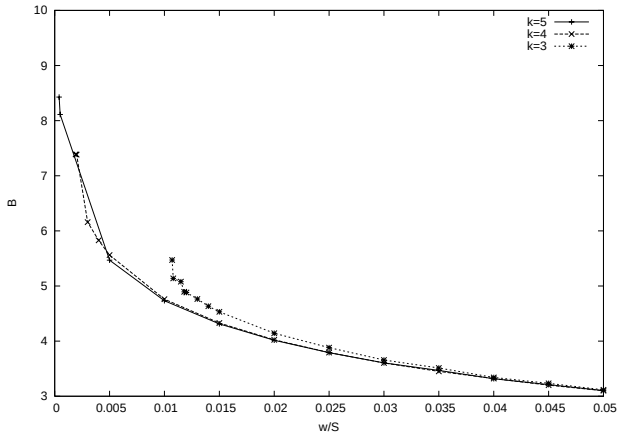

Figura 5.20: GEBB-LBU (d=2,n=100)

0,0015, 0,0006 e 0,0004 para valores de n iguais a 20, 50 e 100, respectivamente.

Para $n = 20$, ilustrado pela Figura 5.18, já se observa uma melhora considerável no desempenho em todas as curvas do gráfico, em relação à segmentação $n = 5$ (Figura 5.17). A melhora na latência, por exemplo, de 0,033 para 0,0061 mediante uma restrição $k = 3$ (redução de um tempo de quase quatro minutos para 44 segundos, considerando um vídeo de duas horas) e de 0,015 para 0,001 quando $k = 5$ (respectivamente, 1m48s para 7,2s).

Para $n = 100$ (Figura 5.20), percebe-se um cruzamento nas curvas de $k = 4$ e $k = 5$ entre as latências de 0,0005 e 0,005, decorrentes da falta de solução por algoritmos genéticos. Nota-se também o ganho pouco expressivo em relação à segmntação $n = 50$ (Figura 5.19). A melhor latência encontrada para $n = 100$ é de 0,0019, um ganho de

0,0002 em relação à melhor solução para $n = 50$. O ganho de largura de banda do servidor também não é expressivo, mas aumenta conforme a latência diminui. Para $w/S = 0,05$, por exemplo, há um ganho na banda passante de $3,146b$ para $3,1b$. Já quando $w/S = 0,004$, a banda passante melhora de $6,029b$ para $5,828b$, um ganho de $0,2b$ em contraste com o ganho de $0,0046b$ para a latência maior. Vê-se aqui o mesmo fenômeno ocorrido com um conjunto de canais (Figura 5.16), onde o ganho inicialmente alto da largura de banda diminui drasticamente com o aumento da segmentação.

Outro aspecto ilustrado pelas figuras é a sobreposição das curvas em algumas faixas de valores de w/S, até que nas curvas onde as restrições são mais fortes ($k = 3$ e $k = 4$) há um desgarramento e então um aumento repentino na demanda de largura de banda. Isto ocorre com maior freqüência para maiores segmentações.

O aumento da segmentação é um fator claro para a redução da latência, pois com a segmentação diminui-se o tamanho do primeiro segmento e conseqüentemente da latência de exibição. Além disso, reduzindo-se a latência de exibição surge um efeito na redução da demanda por banda passante do servidor para a mesma latência. Como exemplo, pode-se comparar um caso na Figura 5.20 onde, para $k = 4$ e $w/S = 0,025$ é necessária uma demanda de largura de banda de $4b$. Para uma segmentação igual a 100 (Figura 5.20), com a mesma largura de banda do servidor, obtém-se uma latência igual a 0,02, o equivalente a um ganho direto de 0,005. Pode-se também afirmar que esta mesma solução para a latência de 0,02 requer $4,39b$ para um problema com 20 segmentos, traduzindo-se num ganho de largura de banda equivalente a $0,39b$.

Uma desvantagem de todos os protocolos adaptados para clientes com limitação de banda passante reside no fato de ser específica para cada limitação. Exemplificando, há um calendário específico para clientes com três canais e outro para quatro canais. Entretanto, pode-se criar faixas de serviços conforme a limitação do usuário. Por exemplo, pode-se fornecer apenas o vídeo, sem operações VCR, para clientes com quatro canais ou mais, e para clientes que possuam cinco ou mais canais oferecer serviços de VCR através de canais contingentes [15].

A Figura 5.21 ilustra a influência exercida pelo parâmetro d no protocolo GEBB-LBU. O gráfico mostra melhorias na latência sem prejudicar na demanda de largura de banda para uma dimensão três; para quatro conjuntos de canais, percebe-se uma grande melhoria na utilização da banda passante, com exceção à faixa entre 0,0020 e 0,0035 (oscilação esta devida à geração de soluções por AG). Já para $d = 5$, nota-se uma queda de rendimento excepcional referente à largura de banda, uma vez que as soluções com esta restrição requeiram em média entre $0,5b$ e b a mais do que as soluções utilizando-se $d = 2$ ou $d = 3$,

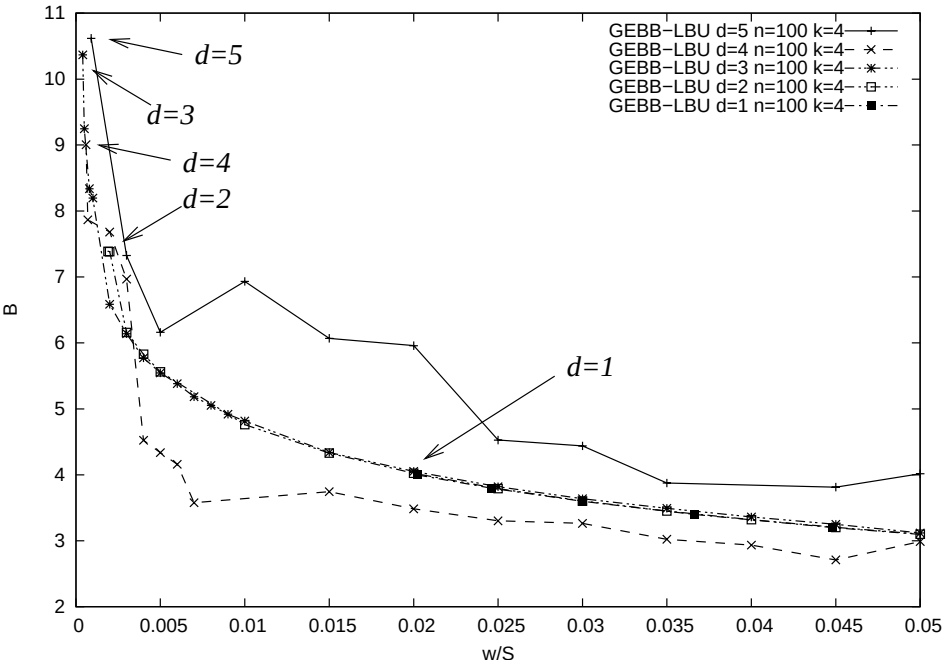

Figura 5.21: Influência do parâmetro d no GEBB-LBU

chegando a uma diferença de mais de $2b$ para uma latência de 1% ($7b$ para a solução de $d = 5$ e $4,8b$ quando $d = 3$). Embora tenha conseguido boas soluções de tempo de espera, não houve uma melhoria de latência em relação aos outros valores de d.

As restrições impostas por problemas de PDPH-LBU e GEBB-LBU com dimensões grandes acabam trazendo problemas durante a fase de geração da população. Uma vez que esta é totalmente aleatória, e em algumas situações não se consegue gerar toda a população para execução do restante do algoritmo. É o que acontece, no exemplo anterior, onde soluções com $d = 4$ aparecem fora da tendência da função, cujas soluções inferiores a soluções com menos de quatro conjuntos de canais (na figura, uma diferença de $0,3b$ para um w/S igual a 0,002 e $0,8b$ para uma fração 0,003).

Outro possível fator para a dificuldade na obtenção de soluções para altas dimensões pode estar na granularidade das soluções. Neste trabalho, utiliza-se valores mínimos para segmentos uma fração de 0,01% do tamanho total de um vídeo, o equivalente a 0,72s para um vídeo de duas horas. Considerando um fluxo MPEG-2 de 6Mbps, o segmento passa a ter um tamanho mínimo de 4,32Mb, o que indica a possibilidade de utilização de uma granularidade ainda maior.

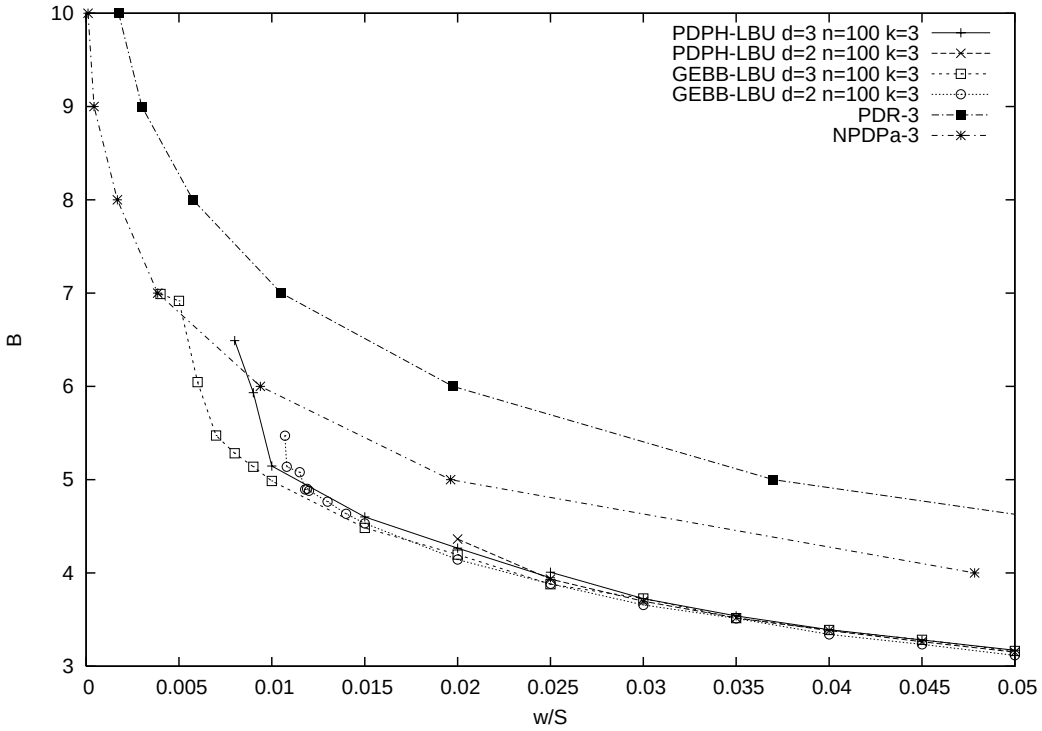

Figura 5.22: Comparação entre PDPH-LBU, GEBB-LBU, PDR-3 e NPDPa-3

5.7 Comparação entre Protocolos com Limitação de Banda Passante dos Usuários

Nesta seção é realizada uma comparação entre os protocolos usuários que possuem limitação na largura de banda. As bases consideradas para comparação são a largura de banda demandanda pelo servidor e a latência mínima obtida. Dois protocolos diferentes somente são comparados quando possuírem a mesma limitação de banda passante do usuário.

Na Figura 5.22, uma comparação dos protocolos otimamente estruturados com outros protocolos que utilizam limitação de largura de banda dos usuário, PDR-3 e NPDPa-3, ambos limitados a uma recepção máxima de três canais.

A versão do GEBB-LBU com um conjunto de canais teve como sua melhor fração de latência o valor 0,0550, o que a excluiu do gráfico, mas pode-se considerar uma marca razoável levando-se em conta que esta versão não possui atraso na recepção do cliente. O mesmo ocorre com a versão do PDPH-LBU, cuja melhor latência obtida foi de 0,0582.

A versão do GEBB-LBU com dois conjuntos de canais obteve um resultado bem mais expressivo que a versão com apenas um canal: 0,0107, ou um minuto e 17 segundos para

um vídeo de duas horas. O ganho do PDPH-LBU também é bem expressivo, obtendo uma latência igual a 0,02.

Já com a versão do GEBB-LBU com três conjuntos de canais ($d = 3$) obteve-se um ganho de latência em relação à versão com dois conjuntos. Para resultados na faixa de latência acima de 0,01, obteve-se em média uma demanda de largura de banda aproximadamente $1,27b$ a menos que o PDR-3 e $0,7b$ a menos que o NPDPa-3. O desempenho do PDPH-LBU é ligeiramente pior, como se pode ver no gráfico.

Para valores de latência entre 0,01 e 0,005 há uma queda de desempenho tanto no PDPH-LBU quanto no GEBB-LBU, e para valores menores 0,005 (o equivalente a 36 segundos para um filme de duas horas) há uma clara vantagem dos protocolos PDR e NPDPa limitados, uma vez que somente estes protocolos conseguem apresentar solução factível. Cabe salientar, entretanto, o aumento de complexidade para os protocolos PDR e NPDPa limitados, assim como para todos da família Pagode, à medida em que a fração do tempo de espera se reduz. Para reduzir o tempo de espera em um protocolo desta família, deve-se obrigatoriamente aumentar o número de segmentos, aumentando assim a complexidade de gerenciamento dos mesmos. Isto não acontece com o PDPH-LBU nem tampouco com o GEBB-LBU, onde a limitação de complexidade é feita através da limitação do parâmetro n_{max}.

O gráfico da Figura 5.23 ilustra a mesma comparação que o gráfico anterior, agora com uma limitação na largura de banda do cliente de quatro canais, onde pode-se notar uma melhoria no desempenho do protocolo GEBB-LBU. Utilizando somente um conjunto de canais, o GEBB-LBU não consegue uma latência mínima boa, pois não há atrasos na recepção dos conjuntos, sendo 2,5% do vídeo (o equivalente a três minutos de espera para um vídeo de duas horas) a menor latência encontrada. Mesmo assim, vê-se neste ponto do gráfico a ótima demanda de largura de banda do servidor, apenas $4b$, aproximadamente b e $1,8b$ a menos que a banda requerida pelo servidor nos protocolos NPDPa-4 e PDR-4, respectivamente. Para um tempo de espera equivalente a 1% (72 segundos para um vídeo de duas horas), GEBB-LBU requer $4,98b$ contra $6b$ do NPDPa-4 e $7b$ do PDR-4. O PDPH-LBU apresenta resultado semelhante, demandando apenas pouca quantidade de banda passante adicional, e demonstrou grande avanço na latência em comparação com a versão limitada a $k = 3$.

O caso do GEBB-LBU com dois conjuntos de canais não foi incluído no gráfico a fim de imprimir maior legibilidade ao mesmo, uma vez que seus resultados são intermediários entre o caso com um e três canais, como ilustrado na Figura 5.21.

Já no caso do GEBB-LBU com três conjuntos de canais, observa-se uma melhora

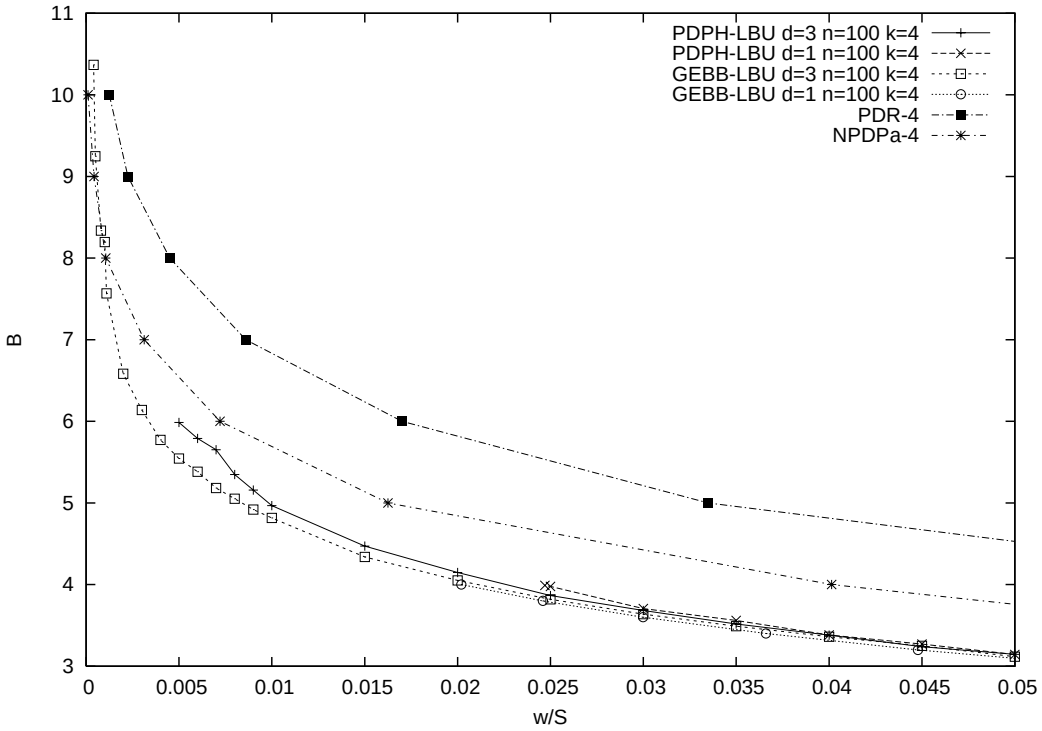

Figura 5.23: Comparação entre PDPH-LBU, GEBB-LBU, PDR-4 e NPDPa-4

sensível no tempo de espera obtido. A menor latência obtida com um resultado superior ao NPDPa-4 é uma fração de 0,0011 (requerendo uma largura de banda de $7,568b$) do vídeo utilizando 100 segmentos, o que equivale a um tempo de espera de 7,92 segundos para um vídeo de duas horas. Este valor é bem satisfatório, perfazendo um bom equilíbrio entre complexidade de gerenciamento, latência e utilização de largura de banda. O protocolo PDPH-LBU também apresentou bom desempenho até a latência de 0,005, mas mesmo assim demandando banda do servidor aproximadamente $2b$ a menos que o PDR-4 e $0,7b$ a menos que o NPDPa-4.

Comparando apenas os protocolos PDPH-LBU e o GEBB-LBU, nota-se uma clara vantagem do GEBB-LBU sobre o PDPH-LBU, obtendo sempre uma demanda por largura de banda do servidor um pouco menor (aumentando a diferença conforme a diminuição da fração w/S). A vantagem do GEBB-LBU na latência obtida é percebida mais claramente, como no caso da utilização de três conjuntos de canais. Enquanto o GEBB-LBU consegue obter um expressivo tempo de espera de 0,04%, contra 0,5% obtido pelo PDPH-LBU. Para apenas um conjunto de canais, o GEBB-LBU obtém uma latência igual a 0,0202, enquanto o PDPH-LBU consegue no máximo 0,0247, o equivalente a meio minuto de diferença em

um vídeo de duas horas.

A diferença torna-se mais evidente mediante o fato de se utilizar a limitação de segmentação $n_{max} = n = 100$. Uma vez que o PDPH-LBU usa este parâmetro por conjunto de canais e o GEBB-LBU, como segmentação máxima independente do número de conjunto de canais (vide Seção 5.6.2), o que permite ao PDPH. uma segmentação potencial d vezes maior do que no GEBB-LBU.

Esta diferença decorre principalmente devido ao tipo de segmentação utilizado pelos protocolos. Embora ambos sejam protocolo otimamente estruturados, uma segmentação crescente como o do GEBB-LBU é muito mais eficaz que a segmentação igual utilizada pelo PDPH-LBU, pois com o mesmo número de segmentos obtém-se um primeiro segmento de tamanho menor, e com isso a latência de exibição obtida também é menor.

Incluindo os gráficos anteriores, pode-se dizer que protocolos otimamente estruturados se beneficiam enormemente de restrições de clientes mais fracas de largura de banda, principalmente na faixa entre $4b$ e $5b$, enquanto possuem um menor desempenho para restrição de largura de banda de clientes variando entre $k = 2$ e $k = 3$, onde não se obteve uma latência de exibição tão boa quanto para valores de k entre quatro e cinco.

Uma vantagem dos protocolos otimamente estruturados é que pode-se atribuir qualquer valor a k, permitindo assim qualquer restrição do usuário. A metodologia utilizada pelos protocolos da família Pagode utilizam sempre valores de k inteiros, pois estes alocam C canais no servidor de largura de banda b e os clientes recebem dados de até k canais simultâneos, o que nem sempre corresponde à realidade. Além disso, para esta família de protocolos é necessário refazer a alocação de segmentos para os canais a cada nova restrição, o que não ocorre com o GEBB-LBU e PDPH-LBU.

A utilização de mensagens de direitos autorais ou mesmo inserções comerciais durante o tempo de espera do vídeo[35] é uma das alternativas para a ocupação do tempo de espera. Protocolos otimamente estruturados têm nesta abordagem uma facilidade adicional, uma vez que a latência de exibição dos mesmos é sempre fixa.

Contudo, os novos protocolos apresentados também apresentam uma desvantagem. Como todos os outros protocolos com clientes limitados em banda passante, tanto o PDPH-LBU quanto o GEBB-LBU são específicos para um valor único de k. Em uma rede heterogênea, com clientes utilizando diferentes restrições de canais, há a necessidade de se utilizar transmissões exclusivas para cada caso ($k = 2$, $k = 3$, $k = 4$, ...).

5.8 Síntese do Capítulo

Este capítulo abordou duas técnicas de adaptação de protocolos de difusão periódica para possibilitar a utilização de clientes com banda passante limitada.

A primeira técnica consiste no atraso da recepção de canais que ultrapassem o limite kb do usuário, e é particularmente eficaz para protocolos da família do Protocolo de Difusão em Pagode — o fato destes protocolos possuírem segmentos de tamanho igual e canais com mesma largura de banda torna fácil a substituição da recepção de um segmento por outro.

Este trabalho propôs uma nova técnica para adaptação direcionada para protocolos otimamente estruturados. Estes protocolos não tiram vantagem da técnica anterior, uma vez que nem sempre possuem mesma largura de banda (no caso do PDPH) ou mesmos tamanhos de segmento (como no GEBB). Esta nova metodologia baseia-se também no atraso na recepção de canais, e utiliza problemas de otimização para determinar qual a configuração ótima para cada caso, dados o número de segmentos, o tempo de espera requerido e o limite de banda do cliente. Os protocolos estudados indicam que o aumento no número de conjunto de canais traz benefícios até certo ponto. Torna-se desaconselhável a solução de problemas utilizando-se mais de três conjuntos de canais para o PDPH-LBU e quatro para o GEBB-LBU.

O protocolo GEBB-LBU mostrara-se eficiente no uso de largura de banda do servidor, tanto em comparação com o PDPH-LBU quanto com outros protocolos (Protocolo de Difusão Rápida e Novo Protocolo de Difusão em Pagode), principalmente entre tempo de espera de 1% a 5% em relação ao tempo total do vídeo.

O NPDPa, em compensação, atingiu níveis de latência não alcançados pelos protocolos otimamente estruturados para uma limitação de três canais simultâneos. Apesar do PDR sempre obter as mesmas latências que NPDPa (tanto para $k = 3$ como para $k = 4$), a largura de banda demandada pelo protocolo é sempre maior (em torno de $2b$) do que a solução do GEBB-LBU e PDPH-LBU em latências acima de 0,01 para $k = 3$ e $k = 4$ e do que o NPDPa para latências inferiores a 0,01.

Capítulo 6

Conclusão

Vídeo sob Demanda faz parte de uma categoria de serviços interativos que permitem a escolha de um vídeo dentre uma vasta coleção. Como vídeos requerem uma demanda muito grande de largura de banda, técnicas de compartilhamento de fluxos são desenvolvidas para diminuir estes requisitos.

Dentre as técnicas de compartilhamento de fluxo, destacam-se os protocolos de difusão periódica por permitirem maior escalabilidade, uma vez que requerem largura de banda constante, independentemente do número de usuários no sistema. Entretanto, a grande maioria dos protocolos de difusão periódica estudados anteriormente não inclui a possibilidade de restrições de recursos do lado do cliente, dentre as quais pode-se destacar a largura de banda. Extensões à protocolos existentes como o de Difusão Rápida e o de Difusão em Novo Pagode já haviam sido feitas. No entanto, estas extensões só são possíveis em protocolos da família onde os segmentos são divididos em tamanhos iguais e transmitidos a uma mesma largura de banda, impossibilitando seu uso em protocolos otimamente estruturados como o Protocolo de Difusão Poliharmônica e o protocolo *Greedy Equal Bandwidth Broadcasting*-GEBB.

6.1 Contribuições

A principal contribuição do trabalho está na criação de dois novos protocolos, o PDPH-LBU e o GEBB-LBU, que utilizam uma nova abordagem na adaptação de protocolos de difusão periódica para suportar transmissão a clientes que possuam banda passante limitada. Esta adaptação se dá através da formulação de um problema de otimização, cuja resolução é dada utilizando-se algoritmos genéticos como heurística.

Nem sempre a limitação de largura de banda de clientes é um múltiplo exato da taxa de consumo dos vídeos. Neste aspecto, os protocolos PDPH-LBU e GEBB-LBU trazem uma vantagem em relação a outros protocolos com limitação de banda passante como o PDR-k e NPDPa-k, uma vez que os primeiros permitem qualquer valor de k.

6.2 Trabalhos Futuros

Protocolos de difusão periódica com limitação de banda passante em clientes são específicos para um valor único de k. Em uma rede heterogênea, com clientes utilizando diferentes restrições de canais requerem transmissões exclusivas para cada caso ($k = 2$, $k = 3$, $k = 4$, ...). A criação (ou verificação de impossibilidade) de um protocolo onde clientes com diferentes larguras de banda possam utilizar-se do mesmo *scheduling* de transmissão do servidor é um trabalho promissor.

A técnica de pré-carregamento dos prefixos (*prefix caching*) pode ser uma boa alternativa para prover acesso instantâneo a vídeos populares transmitidos por um protocolo de difusão periódica. É interessante também realizar um estudo sobre a integração desta técnica com protocolos com limitação de banda do usuário, como o PDPH-LBU e o GEBB-LBU.

Um protocolo de difusão periódica utilizado em conjunto com os fornecimentos de prefixos por demanda e operações VCR eleva-se a um serviço TVoD. A comparação de tal sistema com outros sistemas baseados em difusão seletiva através de simulações também pode ser realizada.

Referências Bibliográficas

[1] Charu C. Aggarwal, Joel L. Wolf, and Philip S. Yu. A permutation-based pyramid broadcasting scheme for Video-on-Demand systems. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems '96*, pages 118–126, Hiroshima, Japão, Junho de 1996.

[2] Michael K. Bradshaw, Bing Wang, Subhabrata Sen, Lixin Gao, Jim Kurose, Prashant Shenoy, and Don Towsley. Periodic broadcast and patching services - implementation, measurement and analysis in an Internet streaming video testbed. UMass CMPSCI technical report 00-56, Department of Computer Science, University of Massachusetts Amherst, 2000.

[3] Ying Cai and Kien A. Hua. Sharing multicast videos using patching streams. *Multimedia Tools and Applications*, 21(2):125–146, November 2003.

[4] Steven W. Carter, Darrell D. E. Long, and Jehan-François Pâris. Video-on-demand broadcasting protocols. In J. D. Gibson, editor, *Multimedia Communications: Directions and Innovations*, pages 179–189, San Diego, 2000. Academic Press.

[5] Y. S. Chen. Mathematical modeling of empirical laws in computer application: a case of study. *Computers and Mathematics with Applications*, 24(7):77–87, Outubro de 1992.

[6] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest. *Introduction to Algorithms*. Cambridge, MA: MIT Press, 1999.

[7] Nelson L. S. da Fonseca and Roberto de A. Façanha. Integrating batching and piggybacking in video server. In *IEEE Global Telecommunications Conference - GLOBECOM'99*, volume 1a, pages 1334–1338, Agosto de 2000.

[8] Asit Dan, Dinkar Sitaram, and Perwez Shahabuddin. Dynamic batching policies for an on-demand server. *Multimedia Systems*, 4(3):112–121, 1996.

[9] Lawrence Davis, editor. *Handbook of Genetic Algorithms*. London: Thomson, 1996.

[10] D. L. Eager, M. K. Vernon, and J. Zahorjan. Minimizing bandwidth requirements for on-demand delivery. In *Proc. 5th Intl. Workshop on Advances in Multimedia Information Systems, Indian Wells, CA, Estados Unidos*, pages 80–87, 1999.

[11] Roberto A. Façanha and Nelson L. S. Fonseca. The Look-Ahead-Maximize-Batching policy. In *IEEE Global Telecommunications Conference*, pages 354–358, 1999.

[12] François Fluckiger. *Understanding Network Multimedia*. London: Prentice Hall, 1995.

[13] David B. Fogel, editor. *Evolutionary Computation: A toward a new philosophy of machine intelligence*. New York: IEEE Press, 1995.

[14] Nelson L. S. Fonseca. *Wiley Encyclopedia of Telecommunications*, chapter Bandwidth Reduction Techniques for Video Services, pages 1–5. Wiley, 2003.

[15] Nelson L. S. Fonseca and Hana K. Rubinsztejn. Dimensioning the capacity of true video-on-demand servers. Technical Report IC-02-07, Instituto de Computação da Universidade Estadual de Campinas, UNICAMP, Agosto de 2002.

[16] Lixin Gao, Jum Kurose, and Don Towsley. Efficient schemes for broadcasting popular videos. In *Proceedings of NOSSDAV*, Cambridge, Reino Unido, Julho de 1998.

[17] Lixin Gao and Don Towsley. Supplying instantaneous Video-on-Demand services using controlled multicast. In *Proceedings of IEEE Multimedia Computing and Systems*, Florencia, Itália, Junho de 1999.

[18] Lixin Gao, Zhi-Li Zhang, and Don Towsley. Catching and selective catching: Efficient latency reduction techniques for delivering continuous multimedia streams. In *Proceedings of 7th ACM International Conference on Multimedia*, pages 203–206, Orlando, FL, Estados Unidos, Outubro/Novembro de 1999.

[19] Leana Golubchik, John C. S. Lui, and Richard R. Muntz. Adaptive piggybacking: A novel technique for data sharing in video-on-demand storage servers. *Multimedia Systems*, 4(3):140–155, Julho de 1996.

[20] Yang Guo, Subhabrata Sen, and Don Towsley. Prefix caching assisted Periodic broadcast: Framework and techniques to support streaming for popular videos. UMass CMPSCI technical report 01-22, Department of Computer Science, University of Massachusetts Amherst, 2000.

[21] Ailan Hu. Video-on-Demand broadcasting protocols: A comprehensive study. In *Proceedings of the IEEE Infocom 2001 Conference*, Anchorage, Alaska EUA, Abril de 2001.

[22] Ailan Hu, Ioanis Nikolaidis, and Peter van Beek. On the design of efficient video-on-demand broadcast schedules. In *Proceedings of the 7th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pages 262–269, College Park, Maryland, Estados Unidos, Outubro de 1999.

[23] Kien A. Hua, Ying Cai, and Simon Sheu. Patching: A multicast technique for true video-on-demand services. In *in Proceedings of 6th ACM International Multimedia Conference*, pages 191–200, Bristol, UK, Setembro de 1998.

[24] Kien A. Hua and Simon Sheu. Skyscraper broadcasting: A new broadcasting scheme for metropolitan Video-on-Demand systems. In *Proceedings of the ACM Special Interest Group on Data Communications Conference (SIGCOMM '97)*, pages 89–100, Cannes, França, 1997.

[25] Li-Shen Juhn and Li-Ming Tseng. Harmonic broadcasting for video-on-demand service. *IEEE Transactions on Broadcasting*, 43(3):268–271, Setembro de 1997.

[26] Li-Shen Juhn and Li-Ming Tseng. Fast data broadcasting and receiving scheme for popular video service. *IEEE Transactions on Broadcasting*, 44(2):100–105, Março de 1998.

[27] John R. Koza, Forrest H. Bennett III, David Andre, and Martin S. Keane, editors. *Genetic Programming III: Darwinian invention and problem solving*. San Francisco: Morgan Kaufmann, 1999.

[28] Scott R. Ladd. *Genetic Algorithms In C++*. New York: M&T Books, 1996.

[29] Maria Eva Lijding, Sape Mullender, and Pierre Jansen. A comprehensive model of tertiary-storage jukeboxes. CTIT technical report series 02-41, University of Twente, Holanda, 2002.

[30] Huadong Ma and Kang G. Shin. Multicast video-on-demand services. *ACM SIGCOMM Computer Communication Review*, 32(1):31–43, Janeiro de 2002.

[31] Jean-Paul Nussbaumer and Frank Schaffa. Capacity analysis of catv for on-demand multimedia distribution. In *Proceedings of the First IASTED/ISMM International Conference on Distributed Multimedia Systems and Applications*, Honolulu, Hawaii, Agosto de 1994.

[32] Jehan-François Pâris. A fixed-delay broadcasting protocol for video-on-demand. In *Proceedings of the 10th International Conference on Computer Communications and Networks (ICCCN'01)*, pages 418–423, Scottsdale, AZ, Estados Unidos, 15 a 17 de Outubro de 1999.

[33] Jehan-François Pâris. A simple low-bandwidth broadcasting protocol for video-on-demand. In *Proceedings of the 8th International Conference on Computer Communications and Networks (IC3N'99)*, pages 118–123, Outubro de 1999.

[34] Jehan-François Pâris, Steven W. Carter, and Darrel D. E. Long. A Hybrid broadcasting protocol for video on demand. In *Proceedings of 1999 Multimedia Computing and Networking Conference*, pages 317–326, Janeiro de 1999.

[35] Jehan-François Pâris, Steven W. Carter, and Darrell D. E. Long. A low bandwidth broadcasting protocol for video on demand. In *Proceedings of the 7th International Conference on Computer Communications and Networks (IC3N'98)*, pages 690–697, Lafayett, LA, Estados Unidos, Outubro de 1998.

[36] Jehan-François Pâris, Steven W. Carter, and Darrell D. E. Long. Efficient broadcasting protocols for video on demand. In *Proceedings of the 6th International Symposium on Modeling, Analysis and Simulations of Computer and Telecommunication Systems*, pages 116–117, Santa Clara, CA, Estados Unidos, Junho de 2000.

[37] Jehan-François Pâris, Darrel D. E. Long, and Patrick E. Mantey. Zero-delay broadcasting protocols for video-on-demand. In *Proceedings of 7th ACM International Conference on Multimedia*, pages 189–197, Orlando, FL, Estados Unidos, Outubro/Novembro de 1999.

[38] Jehan-François Pâris and Darrell D. E. Long. Limiting the receiving bandwidth of broadcasting protocols for video-on-demand. In *Proceedings of the Euromedia 2000 Conference*, pages 107–111, Antuérpia, Bélgica, Maio de 2000.

[39] Sridhar Ramesh, Injong Rhee, and Katherine Guo. Multicast with cache (Mcache): An adaptive zero-delay Video-on-Demand service. In *Infocom*, Abril de 2001.

[40] Jennifer Rexford, Subhabrata Sen, and Andrea Basso. A smoothing proxy service for variable-bit-rate streaming video. In *Proceedings of IEEE Global Internet Symposium*, Rio de Janeiro, Brasil, Dezembro de 1999.

[41] Nelson L. S. Fonseca Roberto A. Façanha and P. J. Rezende. The s2 piggybacking policy. *Multi- media Tools and Applications*, 8(3):371–383, Maio de 1999.

[42] Juan J. Sanchez, Victor M. Gulas, Alberto Valderruten, and Javier Mosquera. State of the art and design of vod systems. In *Proceedings of the International Conference on Information Systems Analysis, SCI'00-ISAS'00*, Orlando, FL, Estados Unidos, Junho de 2000.

[43] Subhabrata Sen, Lixin Gao, Jennifer Rexford, and Don Towsley. Optimal patching schemes for efficient multimedia streaming. UMass CMPSCI technical report 99-22, Department of Computer Science, University of Massachusetts Amherst, 1999.

[44] Subhabrata Sen, Lixin Gao, and Don Towsley. Frame-based periodic broadcast and fundamental resource tradeoffs. UMass CMPSCI technical report 99-78, Department of Computer Science, University of Massachusetts Amherst, 1999.

[45] Subhabrata Sen, Jennifer Rexford, and Don Towsley. Proxy prefix caching for multimedia streams. In *Proc. IEEE Infocom'99*, pages 1310–1319, New York, NY, EUA, Março de 1999.

[46] Andrew S. Tanenbaum. *Redes de Computadores*. Tradução da 3a edição. Rio de Janeiro: Editora Campus, 1997.

[47] S. Viswanathan and T. Imilelinski. Metropolitan area video-on-demand service using pyramid broadcasting. *Multimedia Systems*, 4(4):197–208, 1996.

[48] Bing Wang, Subhabrata Sen, Micah Adler, and Don Towsley. Proxy-based distribution of streaming video over unicast/multicast connections. UMass CMPSCI technical report 01-05, Department of Computer Science, University of Massachusetts Amherst, 2001.

[49] Darrell Whitley. An overview of evolutionary algorithms: Practical issues and common pitfalls. *Information and Software Technology*, 43(14):817–831, 2001.

[50] Joel L. Wolf, Philip S. Yu, and Hadas Shachnai. Disk load balancing for video-on-demand systems. *Multimedia Systems*, 5(6):358–370, 1997.

[51] Rogério M. Zafalão, Nelson L. S. Fonseca, and Cid C. Souza. O protocolo polyharmonic broadcasting sujeito a limitação de banda passante. In *Anais do XXI Simpósio Brasileiro de Redes de Computadores - SBRC, SBC*, volume I, pages 397–410, Natal, RN, Maio de 2003.

[52] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley, 1949.

Apêndice A

Tabela de Símbolos

Símbolo	**Significado**	**Seção**
α	Fator de crescimento do tamanho de segmento	Difusão Piramidal (4.5)
B	Largura de banda utilizada pelo servidor	Divisão de canais (Seção 4.1), PDPH-LBU (5.5) e GEBB-LBU (5.6)
b	Taxa de consumo do vídeo	Capítulos 2 e 5
b^*	Largura de banda ótima	GEBB 4.15
C	Número de canais alocados para um vídeo	Capítulos 2 e 5
$\mathcal{C}$	Conjunto de vídeos não populares	*Batch Min_Idle* (2.3.1)
d	Número de conjuntos de canais	PDPH-LBU (5.5) e GEBB-LBU(5.6)
f	Tamanho de uma fatia de tempo	Protocolo de Difusão em Pagode (4.9) Protocolo de Difusão Harmônica (4.11)
$\mathcal{H}$	Conjunto de vídeos populares	*Batch Min_Idle* 2.3.1
K	Número de canais lógicos multiplexados	Divisão de canais (Seção 4.1)

Tabela A.1: Símbolos utilizados nesta dissertação

Símbolo	**Significado**	**Seção**
m	Número de fragmentos para cada segmento	Difusão Quase-Harmônica (4.13)
m	Número de fatias de tempo de espera; Termo inicial da série harmônica	Difusão Poliharmônica (4.14)
m_{max}	Valor máximo para m	PDPH-LBU (5.5)
n	Número de segmentos	Capítulos 4 e 5
n_{max}	Valor máximo para n	PDPH-LBU (5.5)
P	Fator de multiplexação dos canais	Protocolo de Difusão Piramidal Baseado em Permutações
S	Tamanho do vídeo (em segundos)	Capítulos 4 e 5
S_i	Tamanho do segmento i	Capítulos 4 e 5
V	Número de vídeos transmitidos via difusão i	Pré-carregamento parcial de segmentos (4.16)
W	Tamanho da Janela de Intervalo	*Batching*
W	Limite de tamanho para um segmento	Protocolo de Difusão Arranha-céu
w	Latência de exibição	Capítulos 4 e 5
z	Menor índice de segmento alocado para um determinado canal	Protocolo de Difusão em Pagode (4.9)

Tabela A.1: Símbolos utilizados nesta dissertação

Apêndice B

Lista de Acrônimos

Acrônimo	**Descrição**
ADSL	*Asymmetric Digital Subscriber Line*
AG	Algoritmos Genéticos
ATM	*Asynchronous Transfer Mode*
BML	*Max Batch with Minimal Loss*
CBR	*Constant Bit Rate*
CHB	*Cautious Harmonic Broadcasting Protocol*
CD-ROM	*Compact Disk - Read Only Memory*
DAVoD	*Dynamically Allocated Video on Demand*
DVD	*Digital Versatile Disc*
FB	*Fast Broadcasting Protocol*
FCFS	*First Come First Served*
GDB	*Greedy Disk-conserving Broadcast*
GEBB	*Greedy Equal Bandwidth Broadcasting Protocol*
HB	*Harmonic Broadcasting Protocol*
HDTV	*High Definition TV*
HFC	*Hybrid Fiber/Coax*
IML	*Min Idle with Minimal Loss*
IMQ	*Min Idle with MQL*

Tabela B.1: Tabela de Acrônimos

Acrônimo	**Descrição**
IMQ	*Min Idle with MQL*
ISP	*Internet Service Providers*
JPEG	*Joint Photographic Experts Group*
LAMB	*Look-Ahead Maximize Batch*
LBU	Limitação de Banda do Usuário
MBQ	*Max Batch with MQL*
MFQL	*Maximum Factored Queue Length*
MoD	*Movie on Demand*
MPEG	*Motion Picture Experts Group*
MPLS	*Multiprotocol Label Switching*
MQL	*Maximum Queue Length*
No-VoD	*No Video on Demand*
NPB	*New Pagoda Broadcasting Protocol*
NPDPa	Novo Protocolo de Difusão em Pagode
NTSC	*National Television System Committee*
NVoD	*Near Video on Demand*
PaB	*Pagoda Broadcasting Protocol*
PB	*Pyramid Broadcasting Protocol*
PBR	*Patching Buffer Reuse*
PDA	Protocolo de Difusão Arranha-céu
PDB	Protocolo de Difusão Balanceada
PDH	Protocolo de Difusão Harmônica
PDHC	Protocolo de Difusão Harmônica Cautelosa
PDP	Protocolo de Difusão Piramidal
PDPa	Protocolo de Difusão em Pagode
PDPH	Protocolo de Difusão Poliharmônica
PDQH	Protocolo de Difusão Quase Harmônica
PDR	Protocolo de Difusão Rápida
PDPBP	Protocolo de Difusão Piramidal Baseado em Permutações
PHB	*Polyharmonic Broadcasting Protocol*
PPB	*Permutation-Based Pyramid Broadcasting*

Tabela B.1: Tabela de Acrônimos

Acrônimo	**Descrição**
PVoD	*Partitioned Video on Demand*
PVR	*Personal Video Recorder*
QHB	*Quasi-Harmonic Broadcasting Protocol*
QoS	*Quality of Service*
RAID	*Redundant Array of Inexpensive Disks*
RAM	*Random Access Memory*
RDSI-FL	Rede Digital de Serviços Integrados de Faixa Larga
SB	*Staggered Broadcasting Protocol*
SkyB	*Skyscraper Broadcasting Protocol*
SONET	*Synchronous Optical Networks*
STB	*Set-Top-Box*
TVoD	*True Video on Demand*
VBR	*Variable Bit Rate*
VCR	*Videocassette Recorder*
VoD	*Video on Demand*

Tabela B.1: Tabela de Acrônimos

Índice Remissivo